J. Print Media Technol. Res.
Vol. II - No. 3 (125-214)
September 2013

ISSN 2223-8905
CODEN JPMTBI
UDC 655

3-2013

Journal of Print and Media Technology Research

Thematic issue: Content technologies

Guest Editor Caj Södergård

Scientific contents

Editor-in-Chief Nils Enlund (Helsinki)
Executive editor Mladen Lovreček (Zagreb)
Published by iarigai The International Association of Research Organizations for the Informaion, Media and Graphic Arts Industries
www.iarigai.org

9 772223 890003

J. Print Media Technol. Res.
2(2013)3, 125-214
September 2013

No. 3-2013

ISSN 2223-8905
CODEN: JPMTBI

Journal of Print and Media Technology Research

A peer-reviewed quarterly

PUBLISHED BY

The International Association of Research Organizations
for the Information, Media and Graphic Arts Industries
Washingtonplatz 1, D-64287 Darmstadt, Germany
http://www.iarigai.org E-mail: journal@iarigai.org

A mission statement

To meet the need for a high quality scientific publishing in its research fields of interest, the International Association of Research Organizations for the Information, Media and Graphic Arts Industries (iarigai) publishes the peer reviewed quarterly Journal of Print and Media Technology Research.

The Journal is fostering multidisciplinary research and scholarly discussion on scientific and technical issues in the field of graphic arts and media communication, thereby advancing scientific research, knowledge creation and industry development. Its aim is to be the leading international scientific periodical in the field, offering publishing opportunities and serving as a forum for knowledge exchange between all those scientist and researchers interested in contributing to or benefiting from research in the related fields.

By regularly publishing peer-reviewed high quality research articles, position papers, survey and case studies, the Journal will consistenly promote original research, networking, international collaboration and the exchange of ideas and know how. Editors will also consider for publication review articles, topical and professional communications, as well as opinions and reflections of interest to the readers. The Journal will also provide multidisciplinary discussion on research issues within the field and on the effects of new scientific and technical development on society, industry and the individual. Thus, it will serve the entire research community, as well as the global graphic arts and media industry.

The Journal will cover fundamental and applied aspects of at least, but not limited to the following fields of research:

Printing technology and related processes

◇ Conventional and special printing
◇ Packaging
◇ Printed fuel cells and other printed functionality
◇ Printing on biomaterials
◇ Textile and fabric printing
◇ Materials science
◇ Process control

Premedia technology and processes

◇ Color management and color reproduction
◇ Image and reproduction quality
◇ Image carriers (physical and virtual)
◇ Workflow management
◇ Content management

Emerging media and future trends

◇ Media industry developments
◇ Developing media communication value system
◇ Online and mobile media development
◇ Cross-media publishing

Social impacts

◇ Environmental issues and sustainability
◇ Consumer perception and media use
◇ Social trends and their impact on media

Submissions to the Journal

Submission details and guidelines for authors can be found on the inside back cover of this issue, as well as downloaded from http:// www.iarigai.org/publications/journal.

Subscriptions

http:// www.iarigai.org/publications/journal/order
or send your request to office@iarigai.org.

✉ Contact the Editorial office: journal@iarigai.org

Journal of Print and Media Technology Research

3-2013

September 2013

Thematic issue

Content technologies

Guest Editor

Caj Södergård

J. Print Media Technol. Res. 2(2013)3, 125-214

Printed in Croatia by Narodne Novine d.d. Zagreb

Contents

Peer reviewed papers

Topicalities
Edited by Raša Urbas

A word from the Guest Editor

Caj Södergård

VTT - Technical Research Centre of Finland, Espoo

E-mail: Caj.Sodergard@vtt.fi

Content technologies provide tools for processing content to be delivered via any media to the target audience. These tools are applied in numerous ways in media production. Research into content technologies is very active and opens new possibilities to improve production efficiency as well as to enhance the user experience and thereby the business value of media products and services.

This thematic issue focuses on several applications of content technologies. All papers address the user, and the ability to objectively measure and predict the responses various content causes in users is a much needed tool for the media professional. An emerging application proposed in this issue helps journalists find interesting topics for articles from the excessive information available on the internet. Another class of applications dealt with here is recommending content to the users. Relevant recommendations motivate the user to visit and spend time on a web service. Recommenders are therefore important in designing attractive - and monetizable - digital services. As a consequence, this technology is found in many services recommending media items such as music, books, television programmes and news articles. The papers on recommenders in this issue cover the three main methods in the field - content-based, knowledge-based and collaborative - and they bring new perspectives to all three. One such novel perspective which has been evaluated in user studies is that of a portable personal profile.

Most of the included papers are outcomes of the Finnish *Next Media* research program (www.nextmedia.fi) of Digile Oy. Next Media has run from 2010 through 2013 with the participation of 57 companies and eight research organisations. The volume of the program has been substantial; annually around 80 person years with half of the work done by companies and half by research partners. The program has three foci: e-reading, personal media day, and hyperlocal. The papers in this issue represent only a small part of the results of Next Media. As an example, during 2012 the program produced 101 reports, most of which are available on the web.

Even if this thematic issue is centred on work done within the Finnish Next Media program, content technologies are of course studied in many other places around the world. The paper by NTNU in Norway presented here is just one example. Computer and information technology departments at universities and research institutes often pursue content related topics ranging from multimedia "big data" analysis to multimodal user interfaces and user experience. In the upcoming EU Horizon 2020 program, "Content technologies and information management" is a major topic covering eight challenges. This will keep the theme for this thematic issue in the forefront of European research during the years to come.

Caj Södergård, guest editor of this issue of JMTR, holds a doctoral degree in Information Technologies from the Helsinki University of Technology. After some years in industry, he has held positions at VTT as researcher, senior researcher, team manager and technology manager. His work has resulted in several patents and products used in the media field. Currently Caj Södergård is Permanent Research Professor in Digital Media Technologies at VTT.

JPMTR 021 | 1310
UDC 070:004.5

Original scientific paper
Received: 2013-06-28
Accepted: 2013-11-18

Media experience as a predictor of future news reading

Simo Järvelä[1], *J. Matias Kivikangas*[1], *Timo Saari*[3], *Niklas Ravaja*[1, 2]

[1] Department of Information and Service Economy
School of Business, Aalto University
P.O. Box 21210, FIN-00076 Aalto, Finland

[2] Department of Social Research and
Helsinki Institute for Information Technology
University of Helsinki, Finland

[3] Department of Pervasive Computing
Tampere University of Technology, Finland

E-mails: simo.jarvela2@aalto.fi
matias.kivikangas@aalto.fi
niklas.ravaja@aalto.fi
timo.s.saari@tut.fi

Abstract

The newspaper medium is forced to evolve in the digital age. In order to transfer the core media experience of newspaper reading to new digital formats, its very nature must be examined. In an experiment with 24 readers of a digital newspaper, responses to seven different news sections (people, city, culture, opinion, business, foreign, sports) were measured with psychophysiological methods and self-reports and the differences in responses to them were examined. These data were then compared to actual reading behavior during a six week follow-up period to investigate how immediate media experience predicts future news reading. It was found that the news sections were differentiated by self-reported emotional responses and other message ratings (e.g., relevance to the self, interestingness, reliability), but not by physiological responses. In addition, both self-reports and physiological responses (facial electromyography and heart rate) predicted news reading during the follow-up period, but the strength or direction of the association varied by news section. Different kinds of emotions predict future reading for different news sections, suggesting that people expect differential emotional experiences from different sections.

Keywords: newspaper, readership, media experience, psychophysiology

1. Introduction

Newspaper is a classic media format currently under transformation into new digital formats. The question arises whether the core media experience is altered in this transformation and, if so, how. The term media experience refers to the equivalent of user experience but in the context of media consumption (Hassenzahl, 2008; Kallenbach, 2009; Beauregard and Corriveau, 2007). The newspaper reading experience is a multifaceted construct (Calder and Malthouse, 2004) and emotions are a central element in it, as they are in any media experience (Ravaja, 2004; Helle, Ravaja and Heikkilä, 2011). We set out to study the relatively little researched connection between the media experience and further news reading behavior. Other aspects of newspaper readership have been studied; such as what consumer demographics read which news sections (e.g., Bearden, Teel and Durand, 1978; Burgoon and Burgoon, 1980) and how the content characteristics in news predict newspaper readership (e.g., McCombs, 1987). In more recent years, online news services and their readership have been studied in order to understand the transition to digital media and the changes in readership that follow. For example, a study in 2004 compared the reading of different sections of two large circulation newspapers' paper versions to that of their online versions (d'Haenens, Jankowski and Heuvelman, 2004) and found differences in reading times both between the two papers and also between the various sections. They also found that previous interest predicted reading time of foreign news and business news in both paper and online versions. In the case of online news, several factors are related to the convenience of the medium which may drive adoption. In a study on news adoption to seek for information on public affairs, it was found that many readers turned to internet news sources instead of traditional media. Online news readers are also likely to pursue their own interests in seeking and selecting news instead of following the cues of news editors and producers (Tewksbury, 2003). Among the different reasons for adopting online news are ease and convenience, availability when you want it, timeliness and immediacy, and speed of news access (Conway, 2001). Salwen (2005) concluded that reasons for using online news include being able to get news at any time, being able to directly go to news of interest, easiness and quickness of keeping up with news, convenience, exposure to interesting news stories while doing something else, being able to get different viewpoints, finding unusual news stories online, and being able to get more news than from conventional sources. A recent extensive study by Nguyen (2010) found that factors reflecting how people integrate online news into their everyday life also explain the success of online news: no cost, multitasking, more news choices, in-depth and background information, 24/7 updates, customization, ability to discuss news with peers, and the existence of different viewpoints. Based on this research, it is evident that online news adoption is driven by factors that are related to many of the properties of the internet as a medium and the capacity of this medium to be adopted and domesticated into personalized uses of news as users see best, fitting it into their everyday life. While digitalization of news services has changed news reading, e.g., by adding related features - and it is vital to understand their effects - at the core of reading is still the media experience while consuming the news content itself.

While some studies regarding news section consumption in either print or online media (e.g., d'Haenens et al., 2004) exist, and news reading and emotional reactions to individual news pieces have been studied before (e.g., Ravaja, et al. 2006), the connection of immediate media experience of news sections - the emotions felt during news reading - and it's connection to further news reading is a less studied topic. Most research methods used provide insight into immediate emotions after the news was read. However, the emotions felt during the reading experience itself and those reported immediately afterwards are not necessarily the same (Ravaja, 2004). Many primitive emotional reactions are not conscious but, for example, self-reporting measures are limited to reactions that are subject to various response biases (Paulhus and Reid, 1991; Robinson and Clore, 2002). Psychophysiological methods (Cacioppo, Tassinary and Berntson, 2000) offer a way to assess these emotional reactions with a high temporal resolution by measuring signals obtained by electrocardiography (ECG) or facial electromyography (fEMG). One of the main advantages of this method is its ability to quantitatively measure physiological responses continuously without interrupting the media experience. In the context of news reading, the method allows objective assessment of emotional components such as arousal and valence (Ravaja, 2004; Larsen and Diener, 1992; Lang, 1995) during the actual news reading. Recent studies have shown that psychophysiological measurements in the laboratory can predict subsequent consumer behavior. In a study by Poels et al. (2012), it was found that the measurable reactions predict, in a straightforward manner, what games are played by the subject during the follow-up period and for how long. Kivikangas et al. (2013) obtained similar results, also showing how psychophysiological methods can be utilized in predicting future behavior; however, they claim that which specific signals have predictive power in each case is highly context dependent.

In the current experiment, we set out to explore whether the media experience of various news sections in a newspaper - such as Sports, Business or Culture - differ from each other and if these reactions predict what news the subjects will read in the weeks following the experiment. Utilizing psychophysiological methods and an array of self-report measures, we tried to obtain data concerning various aspects of the news media experience and examine how news reading consumption habits are connected to immediate reactions during news reading. Whether a participant's own stated preferences (background questionnaires and self-reports after the stimulus) or the actual emotional reactions during news reading (psychophysiological recordings) are more accurate in predicting their consumer behavior was also under scrutiny. In times when newspaper media are going through immense changes and digital formats and delivery channels are challenging the more traditional paper formats, the industry is trying to find ways to convert the traditional layout and product design principles into the digital world. It is of vital importance for the industry to acquire precise knowledge of how their products are experienced by consumers and how that affects their long term readership. With more advanced understanding, the new digital formats can be designed to enhance the core media experience.

2. Methods

2.1 Participants

The participants in our experiment were 30 adults aged between 19 and 51 (M=30.4, SD=8.3) who are regular readers of the Helsingin Sanomat (HS) Digilehti (the digital edition of the largest newspaper in Finland, used in the experiment). Due to technical difficulties, 24 participants (10 female, 14 male) were ultimately analyzed.

2.2 Procedure

The participants filled out a background questionnaire and signed a consent form when arriving at the laboratory. After the attachment of electrodes, the HS Digilehti iPad app was shortly introduced to the participants, after which an eight minute baseline was recorded. The participants were instructed to read either an interesting or non-interesting piece of news from a specific news section of HS Digilehti on the iPad according to on-screen instructions. The participants chose which individual article to read based on a quick glance of the headlines and selected either something that seemed interesting or non-interesting depending on the instructtions. Both interesting and non-interesting pieces were chosen to ensure a broader experience for each section. Seven news sections in total were included in the study: Foreign, City, People, Opinion, Culture, Sports, and Business. Separately for each section, the participants were instructed to twice read an *a priori* interesting news story and twice an *a priori* non-interesting news story, resulting in four different pieces of news being read from each section, giving 28 news stories in total. The participants always read the current day's paper so the news were different for all participants as only one experiment per day was conducted - thus cancelling out the effects of content on the results. After reading each news article, the participants filled out self-reports. The instructions were presented and self-reports were obtainned using Presentation on a PC. The laboratory experiment was followed by a six-week period during which the participants continued their normal news reading behavior. Sanoma News (the publisher of the newspaper) provided the participants' full usage data from that period.

The participants were instructed to read the newspaper as they normally would during the follow-up period to avoid any bias. The physiological signals measured were facial electromyography (fEMG), electrodermal activity (EDA), electrocardiography (ECG), electro-encephalography (EEG), and eye-movements. In the current paper we will only discuss the results of self-reports, fEMG, and ECG as some technical difficulties occurred while recording EDA. EEG and eye-tracking data have not yet been analyzed. The physiological signals were recorded using a BrainVision recorder and the eye-movements using SMI eye-tracking glasses. The self-reports included reports on arousal, valence, and dominance using Self-Assessment Manikins (SAMs; Lang, 1980) and an evaluation of the relevance, objectivity, interestingness, reading thoroughness, and reliability of the news.

2.3 Data collection and pre-processing

Facial electromyography (fEMG) activity was monitorred at three muscle sites, *zygomaticus major* (ZM), *corrugator supercilii* (CS) and *orbicularis oculi* (OO), as suggested by Fridlund and Cacioppo (1986). Electrocardiograms (ECG) were measured using the modified lead III electrode placement, and the R-peaks were detected to provide heart rate. For fEMG, the low cut-off filter was 30 Hz, and the high cut-off filter 430 Hz.

The analysis of the raw data was performed using the BrainVision Analyzer v. 2.0.1. The data were filtered using a 50 Hz Notch filter to remove the electric hum. For each reading session, fEMG and ECG data were averaged over the whole reading time. The pre-stimulus baseline was a 3 s period when the participants read instructions before reading each news article. The fEMG signal was rectified and 50 Hz high cut-off filtered.

2.4 Variables

In the end, we had three groups of predictor variables. The physiological variables were the three fEMG activity indices and the heart rate. For self-reported emotion we used Valence, Arousal, and Dominance SAM scales (Lang, 1980). In addition, we used five single-item questions to measure the more subjective assessments of the media experience: we asked how relevant and interesting the news piece was to the participant (Relevance and Interestingness), how thoroughly they read the news piece (Thoroughness), and how reliable and objective they thought the news piece was (Reliability and Objectivity).

To account for the variation of interestingness between the most and the least interesting news stories within a section, we used the *a priori* interest variable (whether the participants chose the news piece as an interesting or non-interesting piece).

As a predicted variable, we used the sum of seconds during which the participants read an article from the particular section during their six-week follow-up period. The number of seconds was derived from the newspaper's customer data following the participants' registered accounts, with articles open less than three seconds being excluded (as they were probably just indications of the reader browsing the newspaper and skipping the article). We also extracted the individual times an article was read from the particular section, but the results were so similar to those based on the usage seconds that we subsequently report mostly only the results using the time as the predicted variable.

2.5 Data analysis

Mean values of the physiological signals during each of the reading periods were calculated for each participant. A mean of EMG from three preceding seconds (relative to a news article) was also calculated as a baseline to check for carryover effects. Facial EMG data and the usage variables were transformed using natural logarithms to normalize their distributions.

The data were analyzed using the Linear Mixed Models (LMM) procedure with restricted maximum likelihood estimation in SPSS 21. The news item identifier was specified as the repeated variable, with participant as the subject variable. Two different sets of analyses were conducted, one with the predictor variables as predicted to see how sections differ from each other in regard to the predictors, and one with predictors predicting the usage.

Section differences (in the variables subsequently used to predict usage) were analyzed using an LMM with fixed effects specified for the section, instruction (*a priori* interestingness), interaction of section and instructtion, and the 3 s local baseline physiological value, and with the predictor as the dependent variable. A random intercept effect with the participant as the subject was included to account for individual differences. Usage predictions were analyzed using an LMM with fixed effects specified for the section, instruction, predictor variable, interactions of section and predictor and of instruction and predictor, and the 3 s local baseline physiological value, where applicable. The same random intercept was included also here.

3. Results

3.1 Differences between sections

Descriptives for usage are shown in Table 1. It is notable that the differences in reading times between sections were substantial. When comparing different sections, the physiological signals fEMG and heart rate (HR) did not differentiate the sections in any way (see Table 2), regardless of whether the news were pre-selected as interesting or non-interesting (*a priori* interest). However, significant differences between sections was found in the majority of the self-reports. They can be most readily detected in Figures 1 and 2.

Table 1: Descriptives of actual reading times (in seconds) during the six-week follow-up, by section

Section	Min	Max	Mean	SD
Overall	0	28 501	1 937.96	3 890.25
People	0	2 587	469.00	662.33
City	0	14 929	3 318.73	4 276.73
Culture	0	5 938	1 324.18	1 580.66
Opinion	0	13 999	1 775.91	3 696.31
Business	0	16 040	2 300.50	4 179.91
Foreign	60	18 506	2 388.05	3 897.63
Sports	0	28 501	19 89.36	5 879.52

Note: SD = Standard deviation

For SAM Valence, news sections were roughly divided into two. People, City, and Culture sections were reported as more positive, while especially the Business section was reported as less positive (the pairwise difference between the smallest value of the former group, Culture, and Business was 0.59, p = .008). While the Foreign, Opinion, and Sports sections were also reported as less positive, the differences were smaller (pairwise differences of 0.44, 0.32, and 0.22, ps = .049, .154, and .339, respectively). For SAM Arousal, the only significant difference was that the Sports section was reported as less arousing than other sections (pairwise difference to the closest section, Business, was .583, p = .004). As can be seen from Figure 1, all reports for valence and arousal were very close to 5, the mid-point of the 9-point scale used, indicating that the participants re-

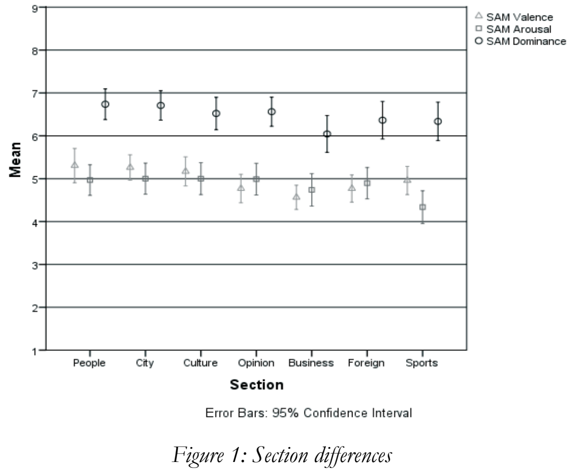


Figure 1: Section differences in self-reported valence, arousal, and dominance

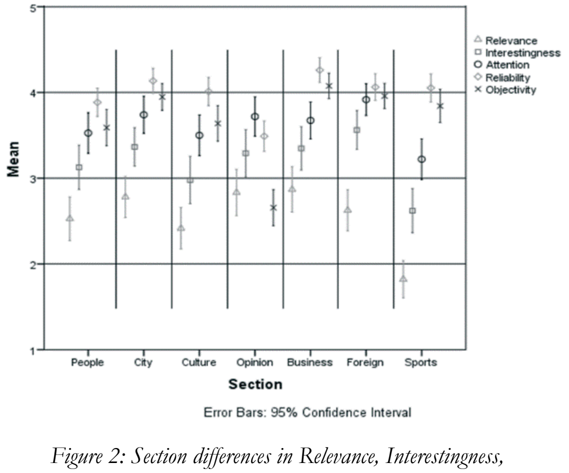


Figure 2: Section differences in Relevance, Interestingness, Thoroughness, Reliability, and Objectivity

Table 2: Overall presence of differences between sections, accounting for whether the particular news article was chosen as interesting or not during the experiment

	Section	*A priori* interest	Section × *a priori* interest
Zygomatic EMG	0.685	1.976	0.473
Corrugator EMG	1.075	1.171	0.509
Orbicularis EMG	0.852	0.183	1.263
Heart rate	0.536	0.471	1.532
SAM Valence	4.192***	110.052***	3.084**
SAM Arousal	3.884***	136.605***	1.026
SAM Dominance	4.792***	17.191***	1.780
Relevance	13.426***	164.394***	3.358**
Interestingness	10.505***	443.507***	1.120
Thoroughness	6.771***	186.411***	0.714
Reliability	16.010***	23.247***	1.252
Objectivity	48.549***	2.049	0.952

Note: Values shown are F-values.
$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

ported almost neutral emotions throughout all sections. For SAM Dominance, however, the ratings were markedly higher, indicating that the participants generally felt being in control during the news reading. Business showed the lowest Dominance rating while People, City, Culture, and Opinion sections showed the highest (pairwise difference between Business and the lowest rating of the four, Culture, was 0.67, *p* = .004). The Sports section was reported as clearly less relevant and interesting than any other section and was read less thoroughly in our sample, pairwise differences ranging from 0.64 to 1.11 for Relevance, from 0.52 to 0.90 for Interestingness (with the exception of the difference to Culture being only 0.23, ns.), all (other) *p*s < .001, and from 0.32 to 0.67 for Thoroughness, all *p*s < .01 (with the exception of Culture, with a difference of 0.24, *p* = .043). Relevance, Interestingness, and Thoroughness showed very similar patterns in other ways as well: The People and Culture sections were also rated relatively low (but not as low as Sports). City, Opinion, and Business were located in the middle, although the differences were not great (at maximum, People and Culture differing from Business in Relevance, *p*s = .013 and .001, and Culture differing from Business in Interestingness, *p* < .001, but Culture differing from Opinion by only *p* = .022 in Thoroughness). This pattern somewhat resembles the actual reading times of the sections (see Table 1). However, although the Foreign section was considered more interesting and was read more thoroughly than others (differing from Culture and People by *p* < .001 and *p* = .005 in Interestingness, and by *p* < .001 and *p* = .003 in Thoroughness), its Relevance was not as high (differing only from Foreign, *p* = .026). Not surprisingly, Reliability and Objectivity of the Opinion section was assessed lower than any other section (all *p*s < .001). Otherwise, Business was rated higher on both scales (comparisons with People, Culture, and Sports > 0.226, *p*s < .016), and People differing also from City and Foreign sections (*p*s < .035). When accounting for *a priori* interest, this affected positively and significantly all other self-reports but Objectivity (see Table 2).

3.2 Predictions

Both a selection of psychophysiological signals and self-reports were used to predict reading time during the six week follow-up period (see Table 3). Perhaps not surprisingly, self-reported Interestingness was the only variable that had general predictive power - i.e., the more interested in a section the participant reported to be, the more he or she read news in that section during the follow-up. The analysis estimated that one point of increase in Interestingness (on a 5-point scale) resulted in reading the news of the section in question for 272.13 seconds (2 ½ minutes) more, constituting a 14 % increase compared to the mean of 1 937.96 seconds. No other variable predicted directly how long the news sections were read. However, when considering the interaction with the sections, the factors Corrugator and Orbicularis EMG, Heart rate, SAM Valence, Arousal, and Dominance, and Relevance, Interestingness and Thoroughness, all predicted actual use significantly. Notably, self-reported Objectivity and Reliability did not predict news reading time, although this may be explained by the small variation in the two variables as we only used news from the largest newspaper of the country.

The main finding is that, when predicting actual news usage, apart from reported interest one must take into account the differences between sections. Although the actual reading times followed the self-reports to some extent, these prediction results show that more information can be captured if one does not simply assume that every section is read in the same way. Instead, when the idiosyncratic differences of the sections are taken into account, a more detailed picture emerges.

Table 3:
Overall effects of predictors when predicting actual reading time

	Predictor	Section × predictor
Zygomatic EMG	1.354	0.432
Corrugator EMG	0.083	4.493***
Orbicularis EMG	0.058	4.156***
Heart rate	3.56	3.419**
SAM Valence	0.025	2.624*
SAM Arousal	0.372	2.35*
SAM Dominance	3.37	4.169***
Relevance	0.191	5.378***
Interestingness	8.288**	2.854*
Thoroughness	2.918	2.525*
Reliability	1.447	1.739
Objectivity	1.372	1.338

Note: Values shown are F-values.
* p < .05, ** p < .01, *** p < .001

Table 4 shows the relative differences in estimates of the interactions by section. The Sports section has been chosen as the reference section (i.e., all the values are in relation to the reference section for that predictor), as its mean reading time was the closest to the overall mean.

From the table it can be seen that for example, for corrugator EMG activity, the Opinion section shows a strong positive association (in relation to the reference section) and the Business section an almost as strong negative association, while for Relevance the associations are reversed (note that while both show negative associations in relation to the reference section, the Opinion section shows a more negative association than the Business section). This demonstrates that a) different sections may have opposite associations with a predictor, and b) different predictors may have opposite patterns of associations in the same sections.

Table 4:
Relative differences in estimates of Section × Predictor interactions for all predictors, by section

Predictor	People	City	Culture	Opinion	Business	Foreign
ZM EMG	-0.23	-0.189	-0.192	-0.332	-0.465	-0.188
CS EMG	0.313	-0.324	0.426	0.57*	-0.493	-0.126
OO EMG	-0.054	-0.837**	0.102	0.062	-0.383	0.548
HR	-0.024	-0.072***	-0.061**	-0.056**	-0.066**	-0.072***
Valence	-0.219*	-0.107	-0.094	0.03	-0.383**	-0.191
Arousal	-0.062	0.069	-0.157	0.092	-0.068	-0.214*
Dominance	0.066	0.259**	-0.126	-0.144	-0.027	0.127
Relevance	-1.12***	-0.753**	-0.818**	-0.944***	-0.553*	-0.841***
Interestingness	-0.433	-0.136	-0.979**	-0.271	-0.689*	-0.981**
Thoroughness	0.74*	0.492	0.864**	0.848**	0.967**	0.505
Reliability	0.513	0.079	0.462	1.107**	0.56	0.082
Objectiveness	-0.379	0.298	-0.433	-0.2	-0.12	0.329

Note: Values shown are the estimates for interaction between the predictor and the section, when the Sports section is used as a reference. Asterisks show the significance of the difference to the reference section. This means that the values should not be considered as absolute values but relative differences between the sections demonstrating the point that different sections are predicted differently with different predictors
* p < .05, ** p < .01, *** p < .001

As a more in-depth example, we have calculated the predicted changes in reading time for some sections (in relation to the reference section) when corrugator EMG activity was used as a predictor and other variables are assumed to be fixed (Figure 3). The test estimates that an increase of 1 (a deepening of the frown) predicts a decrease from 1 347.31 seconds to 770.98 seconds (a drop of 57.2 %) during the six-week follow-up period in the use of the Business section, and use of the Opinion section is predicted to increase from 107.43 seconds to 177.99 seconds (a rise of 60.4 %) in the same time period for the same increase.

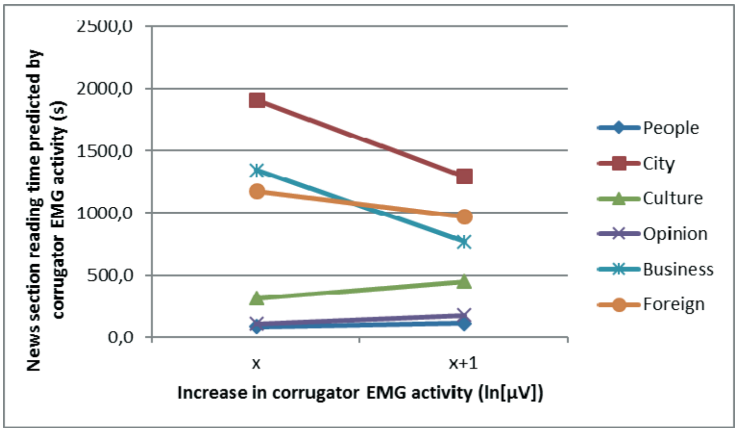


Figure 3: Effects of Section × Predictor interaction for corrugator EMG activity.
The difference between x and x+1 represents a deeper frown measured as an increase of one in corrugator activity unit, and the graphs show how much reading time in the sections change with that increase

4. Discussion

4.1 Interpretations

Our results illustrate how news from different sections elicited different reactions. Our interpretation is that the core affects assessed using psychophysiology were very similar for all sections but that in self-reports, differences could still be reported, partly due to other factors such as reading habits influencing the reporting and simply because in self-reporting replies are commonly relative and participants are capable of always finding some differences when encouraged to do so. For example, the participants reacted differently to news from the Sports section (characterized in self-reports as non-interesting, not relevant and not carefully read) than to news from the Business section (characterized in self-reports as relevant and inducing low valence and dominance). Also, both psychophysiological measurements and self-reports predicted news reading during the follow-up period. However, it was only partly the same measures that differentiated the sections from each other as those that predicted further behavior.

Taken together, the results from self-reports, psychophysiology and usage data can be interpreted as forming two sets of patterns. One set describes the collection of variables that differentiates the sections from each other. The selection of significant variables varies from section to section, thus forming a pattern that describes each section. For example, news from the Opinion section can be described as low valence, high dominance, moderately relevant, interesting and thoroughly read but low on reliability and objectivity, whereas the Business section can be described as low valence and lowest dominance of all sections (though still quite high), also similarly moderate in relevance, interestingness and thoroughness and high on reliability and objectivity. These descriptive traits combined from psychophysiology and self-reports can be thought of as a description of the news section based on the reactions they elicit in the reader.

The second set describes the collection of variables that predict for how long news from that section would be read during the six week follow-up period. Only self-reported interestingness rating predicted all sections - all other variables form an individual pattern for each section. For example, higher negative emotion (CS fEMG activity), higher attention (often associated with the lower HR activity, we found) and high reliability and thoroughness predict the reading of the Opinions section. This finding supports the notion that the ability of psychophysiological methods to predict further behavior is notable, but it is highly context dependent which signals are reactive to the relevant emotional reactions in each case. We have chosen not go through all these individual profiles in depth here, as they are gathered from one local newspaper only and presumably rather idiosyncratic. Instead, we emphasize the general principles present in our findings.

It is notable that, for all news sections, these two patterns - the descriptive and the predictive - are distinct. In other words, the traits that best describe and differentiate news sections are only partly the same ones that predict reading times in real life. Our interpretation is two-fold: firstly, this supports the findings of a previous study regarding the predictive capabilities of psycho-physiological methods (cf., Kivikangas et al., 2013) where it was found that which physiological signals will predict further behavior is context dependent and no single measure is able to consistently do so, regardless of the media content consumed. This emerging pattern also emphasizes that, while sections elicit manifold reactions that have significant differences between each other, not all of those reactions are relevant when it comes to actual consumer behavior. A certain section might elicit certain kinds of reactions but it is another set of reactions that is linked to actual reading habits of the users. In theory, it might be that an optimally designed newspaper could emphasize those features that predict reading more and minimize others. However, as evident from the complexity of the variables and the patterns they form, in practice, this is nearly impossible. Secondly, this separation into two different sets of patterns reflects other relevant variables that were not measured. Reading habits, cultural value statements, etc., are all reflected in the actual reading times. For example, Business section news elicit negative emotions and stress, but are evidently read out of habit and because people feel it necessary to read them regardless of the emotional impact they have.

4.2 Limitations

Our results only shed light on a limited part of the newspaper media experience and more extensive studies should be conducted in order to cover more ground so that a more holistic picture could be formed of how various forms of newspaper media is reacted to and how those reactions are in connection to further consumer behavior. In this study, the consumer behavior was simplified as reading times, something that is easily measurable, but it is clear that consumer behavior and the reading of a newspaper cannot truly be simplified in this way (cf., Malthouse and Calder, 2002). Other studies are required to more extensively study additional reading elements with regard to media experience.

Our study was conducted using content from the largest national newspaper in Finland, but naturally the generalizability of the results is somewhat limited as, e.g., other newspapers may use different divisions into sections, etc. Further studies would be required to affirm the applicability of our results to other newspapers as well. Conducting a similar study with a wider range of demographics and larger sample size would also be beneficial with regard to generalizability. Some results - such as the clearly negative attitudes towards sports - indicate that the sample was at least somewhat biased. However, such biases should not diminish the value of the main findings of this study.

Ultimately a meta-analysis binding all the results from a series of studies together would be in order. Regardless of these short-comings, the results of this explorative study illustrate a clear emerging pattern where news sections elicit distinct reactions. Another separate pattern also emerges that predicts reading behavior; it is our belief that this holds throughout the newspaper media field - both digital and paper - while it is likely that the patterns themselves vary depending on the newspaper in question and the reader demographics.

4.3 Conclusions

The news media field should find our results thought-provoking; in addition to the more obvious results on how sections are different with regard to the reactions they elicit and that the immediate reactions to news reading has an impact on further reading behavior, the fact that these two emerging patterns are not identical is worth delving into. These findings also offer a view on how psychophysiological methods can be utilized in studying news reading behavior and media experience.

The ability of these methods, together with self-report measures, to assess aspects of media experience that are relevant also in long term consumer behavior enables them to be utilized fruitfully throughout the on-going changes in digital media development. On a wider scale, the development of the psychophysiological method and its application to various fields by mapping out the extent of its behavior predicting capabilities is important and this works contributes to that effort.

Acknowledgment

This work received financial support from the Finnish Funding Agency for Technology and Innovation (TEKES) as part of the Next Media research programme, initiated by Digile Ltd. (Finnish Strategic Centre for Science, Technology and Innovation in the field of ICT).

References

Bearden, W. O., Teel, J. E. and Durand, R. M., 1978. Descriptive audience profiles of different daily newspaper sections. *Journal of Applied Communication Research*, 6(1), pp. 31-36

Beauregard, R. and Corriveau, P., 2007. User experience quality: a conceptual framework for goal setting and measurement. *Digital Human Modeling: Lecture Notes in Computer Science*, 4561, pp. 325-332

Burgoon, J. and Burgoon, M., 1980. Predictors of Newspaper Readership. *Journalism Quarterly*, 57(4), pp. 589-596

Cacioppo, J. T., Tassinary, L.G. and Berntson, G. G., 2000. Psychophysiological science. In: J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, eds., 2000. *Handbook of psychophysiology*. New York, NY: Cambridge University Press. pp. 3-26

Calder, B. and Malthouse, E., 2004. Qualitative media measures: Newspaper experiences. *International Journal on Media Management*, 6(1-2), pp. 123-130

Conway, M., 2001. Cybersewers, deserters and includes: An analysis of internet news users and the effect on traditional news media use. In: *Proceedings of the 84th Annual Meeting of the Association for Education in Journalism and Mass Communication.* Washington DC

d'Haenens, L., Jankowski, N. and Heuvelman, A., 2004. News in online and print newspapers: Differences in reader consumption and recall. *New Media and Society*, 6(3), pp. 363-382

Fridlund, A. and Cacioppo, J., 1986. Guidelines for human electromyographic research. *Psychophysiology*, 23(5), pp. 567-589

Hassenzahl, M., 2008. User experience (UX): towards an experiential perspective on product quality. *IHM "08 Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine*, pp. 11-15

Helle, M., Ravaja, N. and Heikkilä, H., 2011. *A theoretical model of media experience and research methods for studying it.* Project report for Next Media - a TIVIT Programme. Helsinki

Kallenbach, J., 2009. Media experience. In: P. Oittinen and H. Saarelma, eds. *Print Media - Principles, Processes and Quality.* Helsinki: Paper Engineers' Association/Paperi ja Puu Oy. pp. 372-410

Lang, P. J., 1980. Behavioral treatment and bio-behavioral assessment: Computer applications. In: J.B. Sidowski, J. H. Johnson and T. A. Williams, eds. *Technology in mental health care delivery systems.* Norwood, NJ: Ablex Publishing. pp. 119-137

Malthouse, E. and Calder, B., 2002. Measuring newspaper readership: A qualitative variable approach. *International Journal on Media Management*, 4(4), pp. 248-260

Malthouse, E., Calder, B. and Eadie, W., 2003. [online] *Conceptualizing and measuring magazine reader experiences*, Available at: http://mediamanagementcenter.sectorlink.org/research/magazineconcept.pdf (Accessed 24 May 2013)

McCombs, M., 1987. Predicting Newspaper Readership from Content Characteristics: A Replication. In: *70th Annual Meeting of the Association for Education in Journalism and Mass Communication* (San Antonio, TX, August 1-4, 1987)

Nguyen, A., 2010. Harnessing the potential of online news: Suggestions from a study on the relationship between online news advantages and its post-adoption consequences. *Journalism*, 11(2), pp. 223-241

Paulhus, D. and Reid, D., 1991. Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, 60(2), pp. 307-317

Ravaja, N., 2004. Contributions of psychophysiology to media research: review and recommendations. *Media Psychology*, 6(2), pp. 193-235

Ravaja, N. et al., 2006. The role of mood in the processing of media messages from a small screen: Effects on subjective and physiological responses. *Media Psychology*, 8(3), pp. 239-265

Robinson, M. and Clore, G., 2002. Belief and feeling: evidence for an accessibility model of emotional self-report. *Psychological bulletin*, 128(6), pp. 934-960

Salwen, M., Garrison, B. and Driscoll, P., 2005. The baseline survey projects: Exploring questions. In: M. Salwen, B. Garrison and P. Driscoll, eds. *Online news and the public.* Mahwah, NJ: Lawrence Erlbaum Associates, pp. 121-145

Tewksbury, D., 2003. What do Americans really want to know? Tracking the behavior of news readers on the Internet. *Journal of Communication*, 53(4), pp. 694-710

JPMTR 022 | 1311
UDC 054:004.4

Original scientific paper
Received: 2013-06-28
Accepted: 2013-11-07

Software Newsroom - an approach to automation of news search and editing

Juhani Huovelin[1], *Oskar Gross*[2], *Otto Solin*[1], *Krister Lindén*[3], *Sami Maisala*[1], *Tero Oittinen*[1],
Hannu Toivonen[2], *Jyrki Niemi*[3], *Miikka Silfverberg*[3]

[1] Division of Geophysics and Astronomy
Department of Physics, University of Helsinki
FIN-00560 Helsinki, Finland

E-mails: juhani.huovelin@helsinki.fi
otto.solin@helsinki.fi
sami.maisala@helsinki.fi
tero.oittinen@helsinki.fi

[2] Department of Computer Science and HIIT
University of Helsinki
FIN-00014 Helsinki, Finland

E-mails: oskar.gross@cs.helsinki.fi
hannu.toivonen@cs.helsinki.fi

[3] Department of Modern Languages
University of Helsinki
FIN-00014 Helsinki, Finland

E-mails: krister.linden@helsinki.fi
jyrki.niemi@helsinki.fi
miikka.silfverberg@helsinki.fi

Abstract

We have developed tools and applied methods for automated identification of potential news from textual data for an automated news search system called Software Newsroom. The purpose of the tools is to analyze data collected from the internet and to identify information that has a high probability of containing new information. The identified information is summarized in order to help understanding the semantic contents of the data, and to assist the news editing process.

It has been demonstrated that words with a certain set of syntactic and semantic properties are effective when building topic models for English. We demonstrate that words with the same properties in Finnish are useful as well. Extracting such words requires knowledge about the special characteristics of the Finnish language, which are taken into account in our analysis. Two different methodological approaches have been applied for the news search. One of the methods is based on topic analysis and it applies Multinomial Principal Component Analysis (MPCA) for topic model creation and data profiling. The second method is based on word association analysis and applies the log-likelihood ratio (LLR). For the topic mining, we have created English and Finnish language corpora from Wikipedia and Finnish corpora from several Finnish news archives and we have used bag-of-words presentations of these corpora as training data for the topic model. We have performed topic analysis experiments with both the training data itself and with arbitrary text parsed from internet sources. The results suggest that the effectiveness of news search strongly depends on the quality of the training data and its linguistic analysis.

In the association analysis, we use a combined methodology for detecting novel word associations in the text. For detecting novel associations we use the background corpus from which we extract common word associations. In parallel, we collect the statistics of word co-occurrences from the documents of interest and search for associations with larger likelyhood in these documents than in the background. We have demonstrated the applicability of these methods for Software Newsroom. The results indicate that the background-foreground model has significant potential in news search. The experiments also indicate great promise in employing background-foreground word associations for other applications.

A combined application of the two methods is planned as well as the application of the methods on social media using a pre-translator of social media language.

Keywords: social media, data mining, topic analysis, machine learning, word associations, linguistic analysis

1. Introduction

The vast amount of open data in the internet provides a yet ineffectively exploited source of potential news. Social media and blogs have become an increasingly useful and important source of information for news agencies and media houses. In addition to the news collected, edited and reported by traditional means, i.e., by news agencies, the information in a news-room consists of different types of user inputs. In the social media there is a large amount of user comments and reactions triggered by news stories. Also, fresh article manuscripts and other types of material can be produced by basically anyone by submitting the information to the internet. As a means of collecting news, this material is already in use by commercial media companies, especially in a hyperlocal media context (e.g., newspapers that discuss local issues).

While this editorial strategy is considerably more advanced than the way news were produced a decade ago, the work still includes manual work that could be automated and the use of open data available in the internet is usually very inefficient. It also does not make much sense to engage humans for browsing internet data, a job that can be done much more efficiently and tirelessly by a machine.

Thus, intelligent computer algorithms that monitor internet data and hunt for anomalies and changes are becoming an increasingly exploited means of news and trend detection. Other applications for the same methodologies are public opinion analysis and forecasting the results of elections. Examples of even more advanced intelligence in prediction would be calls to events, which can be predecessors of demonstrations or even an uprising, and indication of a meeting between high level politicians based on their plans to travel to the same place at the same time.

The same methods, when combined with fusion of heterogeneous data, can help improving the quality and widening the scope of news by the enrichment of existing news material with relevant background information and other associated material (e.g., history, pictures, digital video material). In principle, using the same methodology it is also possible to follow the discussions raised by published news articles and thus automatically collecting feedback from the audience.

Examples of internet services developed for the above purposes are Esmerk Oasis (Comintelli, 2013) and Meltwater Buzz (Meltwater, 2013). Esmerk Oasis is a web-based market intelligence solution. Its services include customized global business information with the possibility of importing complementary information from other sources as well as sharing and distribution of information across the client organization. Meltwater Buzz is a social media monitoring tool that has capabilities for tracking and analyzing user-generated content on the web. Google has also developed several services that perform similar tasks.

Considering the purpose and goal of an automated news search and analysis process, a baseline approach to analyzing text material and creating a short descripttion of its contents is to simulate the traditional process of news production. The analysis of the material should tell you *what, who, where, and when*? Methodologically, the most challenging task is to find a systematic way of defining the answer to the question *what*, since it includes the need to recognize and unambiguously describe an unlimited range of *topics,* not just individual words. A topic is usually defined as "a set of news stories that are strongly related by some seminal real-world event", and an event is defined as "something (non-trivial) happening in a certain place at a certain time" (Allan, 2002). As an example, the recent meteorite impact in Chelyabinsk was the event that triggered the asteroid impact, natural catastrophes, and doomsday topic. All stories that discuss the observations, consequences, witnesses, probabilities and frequency of such events etc., are part of the topic.

The answers to the other questions*, who, where* and *when,* can be traced by searching named entities and various time tags and information. In practical application to, e.g., social media, however, the latter questions may also pose a significant challenge for an automated approach, since social media language does not obey common rules.

The quality of the language is often very poor, since it may include many local and universal slang words, acronyms and idioms that are known by only a limited local community, and also numerous typing errors.

Blogs are considerably less difficult in this respect, since most of the text in them is in fairly well written standard language.

Methods for event and trend detection and analysis in large textual data include *static and dynamic component models* which are well suited for news search and detection in the internet. Static models are simpler to use and give results that are easier to interpret. A potential disadvantage is that newly emergent trends may remain undetected if the training data for the model is not sufficiently extensive, leading to the model being not generic enough. A dynamic model, on the other hand, is updated continuously in order to keep up with possible emergent topics. Its usage, however, is not as straightforward as that of static models since the emergent trends may be described in terms of dynamic components whose semantics is not yet well understood.

An example of a static component model is Principal Component Analysis (PCA). PCA was invented in 1901 by Pearson (1901). PCA can be performed by eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix. The singular value decomposition of the word count matrix is also cal-led Latent Semantic Indexing (LSI) (Berry, Dumais and O'Brien, 1994; Hofmann, 1999).

We have developed algorithms for automated analysis of text in, e.g., social media, blogs and news data with the aim of identifying "hot" topics that are potential news. We here present the methods and show results of their application using real data.

2. Method

2.1 Combining methods

We apply two different approaches that are combined in order to achieve a clearer recognition of potential news in an arbitrary text under analysis. The first method is topic mining including advanced linguistic analysis for named entity recognition. The topic model is based on Multinomial Principal Component Analysis, MPCA (Kimura, Saito and Uera, 2005; Buntine and Jakulin, 2006). While topics are considered to be different kinds of objects than named entities (e.g., Newman et al., 2006), they can be combined in the creation of a probabilistic topic model. The second approach, association analysis, takes into account the word co-occurrences in the document and uses statistics to look for novel word associations in a set of documents. These associations are used for Software Newsroom applications, such as diverging (association) word clouds and automatic summary generation. The background model calculation uses a method based on the log-likelihood ratio (LLR) (Dunning, 1993). This is described in more detail by Toivonen et al. (2012). By extending the ideas in the latter approach, we propose a method for detecting novel word associations.

2.2 Topic mining

For generating static component models from textual data, we use the statistical generative model called *Multinomial principal component analysis* (MPCA) (Buntine and Jakulin, 2006). MPCA is used to model the data in order to obtain a comprehensive understanding of the contents of the data sources in the form of semantically meaningful components or topics.

In our application, the topic model includes four categories of common words (nouns, verbs, adjectives, adverbs) where the nouns are not named entities, and four categories of named entities (persons, places, organisations, miscellaneous), where the miscellaneous category includes all named entities that do not belong to the other three categories. It has been shown that these eight categories are effective for building topic models for English (e.g., Newman et al., 2006). An important aspect of our research is to verify that the linguistic categories can be identified in a language-independent way. We demonstrate this by extracting the eight categories of words from text in Finnish - a language completely unrelated to English.

Let $\mathbf{D}$ be a $d \times N$ matrix representing the training data (documents) as a "bag-of-words", $\mathbf{M}$ a $d \times K$ matrix of documents represented in terms of topics, and $\mathbf{\Omega}$ a $K \times N$ matrix of topics represented in terms of words, where d is the number of documents in the training corpus, N the size of the vocabulary and K the number of topics ($K << N$). We extract from the training corpus two types of features: Part-of-speech tags (nouns, adjectives, verbs, adverbs) and Named Entities (locations, persons, organizations, miscellaneous). Thus, in our case, the vocabulary words are treated as eight multinomials. The aim is to represent the documents in terms of matrices $\mathbf{M}$ and $\mathbf{\Omega}$ as (Equation 1):

$$\mathbf{D} \approx \mathbf{M} \times \mathbf{\Omega}. \qquad [1]$$

In other words, the data is transformed into a lower dimensional space, where documents are represented in terms of topics. The topics are then represented in terms of words. The matrices $\mathbf{M}$ and $\mathbf{\Omega}$ give the probabilities of topics given a document and words given a topic, respectively.

The process for generating the model with MPCA is as follows (Buntine and Jakulin, 2004).

1. A total of N words are partitioned into K partitions $\mathbf{c} = c_1, c_2, \ldots, c_K$ where $\sum_{k=1}^{K} c_k = N$. N is the size of the vocabulary and K the number of topics. The partitioning is done using a latent proportion vector $\mathbf{m} = (m_1, m_2, \ldots, m_K)$. The vector $\mathbf{m}$ for each document forms the d rows in matrix $\mathbf{M}$.

2. Words are sampled from these partitions according to the multinomial for each topic producing a bag-of-words representation $\mathbf{w}_{k,\cdot} = (w_{k,1}, w_{k,2}, \ldots, w_{k,N})$ for each partition k.

3. Partitions are combined additively to produce the final vocabulary $\mathbf{r} = (r_1, r_2, \ldots, r_N)$ by totaling the corresponding counts in each partition,

   $r_n = \sum_{k=1}^{K} w_{k,n}$

The above process is described by the following probability model (Equations 2),

$$\begin{aligned} &\mathbf{m} \sim \mathit{Dirichlet}(a) \\ &\mathbf{c} \sim \mathit{Multinomial}(\mathbf{m}, N) \qquad [2] \\ &\mathbf{w}_{k,\cdot} \sim \mathit{Multinomial}(\mathbf{\Omega}_{k,\cdot}, c_k) \quad \text{for } k = 1,\ldots,K \end{aligned}$$

Its estimation is done through a Gibbs sampler (Buntine and Jakulin, 2006). In Gibbs sampling, each unobserved variable in the problem is resampled in turn according to its conditional distribution. Its posterior distribution conditioned on all other variables is computed, and then a new value for the variable using the posterior is sampled. In each cycle of the Gibbs algorithm the last $\mathbf{c}$ for each document is retrieved from storage and then, using a Dirichlet prior for rows of $\mathbf{\Omega}$, the latent component variables $\mathbf{m}$ and $\mathbf{w}$ are sampled. The latent variables are $\mathbf{m}$ and $\mathbf{w}$, whereas $\mathbf{c}$ is derived. As a result we get estimates of the matrices $\mathbf{M}$ and $\mathbf{\Omega}$ in

Equation 1. In the context of an MPCA model, these are estimates of the distribution of documents over topics and topics over words respectively.

There are different ways of estimating topic strengths in a single document given the model created by MPCA. The method applied here is cosine similarity between document vector *d* and topic ω_k as

$$\mathrm{sim}(\mathrm{d}, \omega_\mathrm{k}) = \frac{\mathrm{d} \cdot \omega_\mathrm{k}}{||\mathrm{d}||\,||\omega_\mathrm{k}||}. \qquad [3]$$

A topic model including a desired number of yet unnamed topics is first created by the above method (Equations 1 and 2) using a bag-of-words presentation of the training corpus. This is then ready for application in topic analysis of arbitrary text. The topic analysis includes automated simultaneous identification of the topic, person, place, organization, and event in an arbitrary blog article, a discussion thread in social media or an RSS feed, etc. This is done by statistical comparison, or projection (Equation 3), of the new text against the topic model. By tracking the history of the frequency of occurrence of similar stories (which belong to the same topic, i.e., resemble each other), the software can identify the trend of a topic. A statistically significant deviation from the trend in a short time period gives a hint that the source texts that caused this deviation may include a news candidate. In the present analysis, we use Gaussian statistics and the criteria for significant deviation is 3σ. This applies to generic topics, but for words and events that are generally interesting from a newsroom perspective, such as VIPs, accidents, crimes, and natural disasters, all occurrences are tagged as being potential news topics. The method is, on a general level, similar to the approach of Newman et al. (2006) but it includes advanced features developed for practical applicability in a newsroom environment.

2.3 Linguistic analysis for named-entity recognition

2.3.1 English vs. Finnish words and named entities

When adapting topic identification from one language to another, it is necessary to be aware of what units of the language have been chosen and how similar units can be identified in another language. All language analysis methods do not produce the same output granularity. In the following, we outline the units that have been found effective in English and how corresponding units can be identified in Finnish to highlight some of the essentials that need to be considered when choosing linguistic analysis software to adapt to another language.

The most striking difference between Finnish and English is the number of inflected forms in Finnish. There are roughly 2000 forms for each noun, 6000 for each adjective and 12000 for each verb. The characteristics of these forms and their usage in Finnish has been extensively documented in an online Finnish grammar, "Iso suomen kielioppi" (Hakulinen et al., 2004). It is not possible to only chop off word endings, because changes also take place in the stem when inflectional morphemes are added, e.g., "nojatuoli" [armchair], "nojatuoleja" [armchairs], "nojatuoleissa" [in the armchairs], "nojatuoleissani" [in my armchairs], "nojatuoleissanikin" [also in my armchairs]. In practice, Finnish words can represent expressions that in English are rendered as a phrase, so Finnish needs a morphological analyzer to separate the base form from the endings. As a bonus to the morphological processing, many of the inflectional morphemes that are separated from the base form correspond to stop-words in English.

In addition to gluing inflectional morphemes onto the words, Finnish also has the orthographic convention of writing newly formed compound words without separating spaces, i.e., "nappanahkanojatuoli" [calf-skin armchair]. The English word *armchair* can be seen as a compound as well, but typically a modern *armchair* is not perceived only as a *chair* with *arm*rests, but as something slightly more comfortable, so the *armchair* has a lexicalized meaning of its own. This means that, for newly coined non-lexicalized compounds, it is essential that the morphological analysis separates the non-lexicalized parts in Finnish; otherwise the compositional meaning is lost. Long newly formed compounds also lack predictive power since they are rare by definition whereas the compound parts may give essential clues to the topic of the narrative. It should be noted that a Finnish writer could also choose to write "nappanahkainen nojatuoli" [calf-skin armchair], and with the increased influence of English, this convention is perceived as more readable.

The structure of named-entities, i.e., places, organizations, persons and other names, follows the conventions mentioned for regular words. In particular, place names tend to be written in one or two words at most because they are of older origin. Person names have a similar structure as in English with given name and surname. However, long organization names tend to be formulated as multi-word expressions following newer writing tendencies.

2.3.2 Named-entity recognition in Finnish

For named-entity recognition in many languages it is possible to do string matching directly on the surface forms in written text. In Finnish, we need more in-depth morphological processing to deal with the inflections and the compound words. For out-of-vocabulary words, we also need guessers. To cope with morphological ambiguity, we need a tagger before we can apply named-entity recognition.

Language technological applications for agglutinating languages such as Finnish, benefit greatly from high coverage morphological analyzers providing word forms with their morphological analyses, e.g.,

"nojatuole+i+ssa+ni+kin : nojatuoli *Noun Plural 'In' 'My' 'Also'*" [also in my armchairs].

However, morphological analysis makes applications dependent on the coverage of the morphological analyzer. Building a high coverage morphological analyzer (with an accuracy of over 95%) is a substantial task and, even with a high-coverage analyzer, domain-specific vocabulary presents a challenge. Therefore, accurate methods for dealing with out-of-vocabulary words are needed.

With the Helsinki Finite-State Transducer (HFST) tools (Lindén et al., 2011), it is possible to use an existing morphological analyzer for constructing a morphological guesser based on word suffixes. Suffix based guessing is sufficient for many agglutinating languages such as Finnish (Lindén and Pirinen, 2009), where most inflection and derivation is marked using suffixes. Even if a word is not recognized by the morphological analyzer, the analyzer is likely to recognize some words which inflect similarly as the unknown word. These can be used for guessing the inflection of the unknown word.

Guessing of an unknown word such as "twiitin" (the genitive form of "twiitti", tweet, in Finnish) is based on finding recognized word forms like "sviitin" (genitive form of "sviitti" hotel suite in Finnish), that have long suffixes such as "-iitin", which match the suffixes of the unrecognized word. The longer the common suffix, the likelier it is that the unrecognized word has the same inflection as the known word. The guesser will output morphological analyses for "twiitin" in order of likelyhood.

A morphological reading is not always unique without context, e.g., "alusta" can be an inflected form of "alku" [beginning], "alunen" [plate], "alustaa" [found] or "alus" [ship]. To choose between the readings in context it is possible to use, e.g., an hidden Markov model (HMM) which is essentially a weighted finite-state model. Finite-state transducers and automata can more generally be used for expressing linguistically relevant phenomena for tagging and parsing as regular string sets, demonstrated by parsing systems like Constraint Grammar (Karlsson, 1990) which utilizes finite-state constraints. Weighted machines offer the added benefit of expressing phenomena as fuzzy sets in a compact way.

Using tagged input, a named entity recognizer (NER) for Finnish marks names in a text, typically with information on the type of the name (Nadeau and Sekine, 2007). Major types of names include persons, locations, organizations and events. NER tools often also recognize temporal and numeric expressions. NER tools typically use gazetteers, lists of known names, to ensure that high-frequency names are recognized with the correct type. For Finnish, the gazetteer is included in the morphological analyzer because names inflect. In addition, names and their types can be recognized based on internal evidence, i.e., the structure of the name itself (e.g., ACME Inc., where *Inc.* indicates that ACME denotes a company), or based on external evidence, i.e., the context of the name (e.g., *works for* ACME; ACME *hired a new CEO*) (MacDonald, 1996).

2.4 Association analysis

2.4.1 Extracting word associations

One of the goals of the Software Newsroom is to give an overview of popular topics discussed in the internet communities. This gives journalists an opportunity to react to these topics on a short notice. In the Software Newsroom, word association analysis is used for detecting novelty in the contents of a given set of documents. For instance, consider a web forum where people discuss about different topics, e.g., fashion, technology, politics, economics, computer games, etc. As an example, consider that a new smartphone *SoftSmart* has a feature which automatically disables GPS when you are indoors. It turns out that it has a bug, and in some very specific cases (e.g., for instance when you are on the top floor of a building) it starts to drain your battery because the signal strength is varying. It is reasonable to believe that many *SoftSmart* users will go to web forums and start discussing about the problems. Even more, it might turn out that there is an easy fix available and this is posted somewhere to the forum. The problem is, that there are thousands of similar problems being discussed all over the world, so it is not feasible for a technology journalist to monitor all the forums.

If we could automatically detect this as a trendy topic, then this information would be invaluable for a technology journalist, as she/he could then learn more about this and write a news story. From the language analysis point of view, the text written by people in web forums and other web communities introduce problems - the text contains slang, typing errors, words from different languages, etc. These aspects add another goal for the association analysis - our goal is to develop a method which is not fixed to any specific vocabulary. Our idea is to analyze the associations between words and to look for such associations which are novel with regards to other documents.

Considering the *SoftSmart* example, there are words which co-occur in sentences but the association between them is most probably very common, such as *SoftSmart - battery, battery - drain, SoftSmart - GPS,* etc. For the *SoftSmart* case, the words for which the association is rather specific could be *battery - floor, floor - drain, battery - top, Softsmart - floor* and so on. In association analysis, our goal is to automatically detect the latter ones. Note, that the association itself might be surprising, though it is between very common words, like 'battery' and 'floor'.

Finding associations between concepts which can be represented as sets of items is a very much studied area which originates from the idea of finding correlations in market basket data (Agrawal et al., 1993). The bag-of-words model of representing documents as sets of unordered words is a common concept in information retrieval (Harris, 1954; Salton, 1993). Often, the bag-of-words model is used together with the tf-idf measure that measures word specificity with respect to the document corpus (Salton, 1993).

Analysing word associations in document is not a new idea. There are various word association measures available - the log-likelihood ratio test (Dunning, 1993), the chi-squared test, Latent Semantic Indexing (Dumais et al., 1988), pointwise mutual information (Church and Hanks, 1990), Latent Dirichlet Allocation (Blei et al., 2003), etc. There is also a method for pairs, which is inspired by tf-idf, called tpf-idf-tpu which is a combination of using term pair frequency, its inverse document frequency and the term pair uncorrelation for determining the specific pairs of a document (Hynönen et al., 2012).

In this paper, we present a method for analyzing and representing documents on the word association level. We use the log-likelihood ratio as the basis for our method. As mentioned before, finding associations between documents is a very common concept and the main goal for all the methods is to discover statistically strong associations between words. In some instances we are interested in such associations that are specific to a certain set of document. For instance, consider a set of documents about the singer Freddie Mercury. Imagine, that we create pairs of all the words which co-occur in the same sentence and the weight is determined by their co-occurrence statistics (e.g., weighted by the log-likelihood ratio test). Now, if we order the pairs decreasingly by association strength, we will most probably obtain pairs such as: 'freddie'-'singer', 'freddie'-'aids', 'freddie'-'bohemian', 'aids'-death', 'aids'-sick' etc. The point here is that some of the associations are important and relevant to the document set (e.g., the first two). On the other hand, the last two associations between words are very common. And this is defines our goal - we are looking for word associations which are specific to a certain set of documents *and* at the same time are uncommon with respect to other documents.

In the following, we introduce methods for extracting word associations that are specific to a set of documents. For this we define two concepts: *background associations*, which are the common associations between words and *foreground associations*, where the weight is higher for associations that are novel with respect to the background associations.

After we have given an overview of the core methods, we will present applications of these models in the Software Newsroom. First we will look at the possible representations of foreground associations and discuss the possible usefulness of explicit graph representations. Then we will provide an idea of diverging word clouds which illustrate word associations rather than frequencies. Finally, we propose a simple, yet intuitive way of generating summaries of a set of documents by using foreground associations.

2.4.2 Background associations

Background associations represent common-sense associations between terms, where the weight depends on the strength of the association. For example, the connection between the words 'car' and 'tire' should be stronger than the connection between 'car' and 'propeller'. In our methodology, these associations are extracted from a corpus of documents, motivated by the observation that co-occurrence of terms tends to imply some semantic relation between them (slightly misleadingly often called semantic similarity). Background associations are calculated by identifying words which co-occur in the same sentence. The strength between the words is calculated using the log-likelihood method (Dunning, 1993; Toivonen et al., 2012). The latter paper describes how the word associations are calculated and also demonstrates the relationship between such associations and relations in WordNet (Miller, 1995).

2.4.3 Foreground associations

In contrast to the common associations in the background, *foreground associations* represent novel associations of a (small) set F of documents called the foreground documents.

However, the background associations do have a central role here: they tell us what is known, so that we can infer what is novel in any given document. The weighting scheme in the foreground also uses the log-likelihood ratio test. However, now we use the background to obtain the expected number of co-occurrences and to see how much the observed number of co-occurrences in the foreground documents F deviates from it. The result of this test gives higher weights to those term pairs that are more frequent in the foreground F than they are in the background, i.e., especially those pairs which have a small likelihood of occurring together in the background.

In our implementation of this idea, the foreground weights are based on the log-likelihood ratio where the alternative model is based on the foreground documents F and the null model on the background corpus C.

Let parameters p_{ij}^{null} be the maximum likelihood parameters for the corpus C, i.e., (Equation 4):

$$\begin{aligned} p_{11}{}^{null} &= p(x \wedge y;\ C) \\ p_{21}{}^{null} &= p(x \wedge \neg y;\ C) \\ p_{12}{}^{null} &= p(\neg x \wedge y;\ C) \\ p_{22}{}^{null} &= p(\neg x \wedge \neg y;\ C) \end{aligned} \qquad [4]$$

where *x* and *y* denote the events that "word x (respectively y) occurs in a randomly chosen sentence (of the given corpus)". For the background associations, these parameters are used as the alternative model, and here they are used as the null model. Set the alternative model parameters p_{ij} in turn to be the maximum likelyhood parameters for the document set F (Equations 5),

$$\begin{aligned} p_{11} &= p(x \wedge y;\ F) \\ p_{21} &= p(x \wedge \neg y;\ F) \\ p_{12} &= p(\neg x \wedge y;\ F) \\ p_{22} &= p(\neg x \wedge \neg y;\ F) \end{aligned} \qquad [5]$$

The log-likelihood ratio (LLR) for the foreground associations is then computed according to Equation 6.

$$LLR(x, y) = -2 \sum_{i=1}^{2} \sum_{j=1}^{2} k_{ij} \log(p_{ij}{}^{null}/p_{ij}) \qquad [6]$$

The foreground association weights are assigned by this LLR function. Using this function, we give higher weight to such associations which are more likely to appear in the foreground and less likely in the back-ground. Note, that the log-likelihood ratio could be also negative. In this case the word association is weaker in the foreground than in the background. In our work we omit associations with negative weights.

2.4.4 Applications

In the following, we present applications in the Software Newsroom that employ the background/foreground associations method. In the first application we describe, the associations in the set of documents are represented as an explicit graph. In the remainder of the subsection we will demonstrate two different Software Newsroom applications - diverging word cloud generation and document summarization. For a single document experiment we will use the English Wikipedia as the background corpus and a story from BBC: "Google tests balloons to beam internet from near space" (Kelion, 2013) as the foreground document we are interested in.

The simplest way of representing the information is by showing the top-k (where k is an integer) word pairs of the news story. In order to show the differences between a standard co-occurrence calculation and our foreground method in we have, in Table 1, presented the top-5 pairs of the Google news story. For comparison, the left column lists the most strongly associated word pairs as measured using standard methods, while the right column lists the top-5 pairs obtained by the foreground method.

The pairs suggest that the foreground method is able to grasp the main associations of the news story better than the classical co-occurrence measures. By this we mean that the associations of the foreground contain more relevant associations, such as 'superpressure' and 'balloons' or 'google' and 'balloons'. Representing associations as a simple list makes them individually easy to understand, but does not give a picture of the network of connections. On the other hand, a graphical representation (Figure 1) of this network may be difficult, especially for novice users. On the other hand, when a user is familiar with such data representation it gives a quick and general view of the data. In our work, the explicit graph is not a favored method for illustrating or representing information. We put more emphasis on designing methods that employ the foreground graph.

Table 1: The top-5 pairs for the BBC news story "Google tests balloons to beam internet from near space". The left column shows pairs calculated using the standard co-occurrence calculation method (log-likelihood ratio); the right column shows the top-5 pairs obtained using the foreground association method

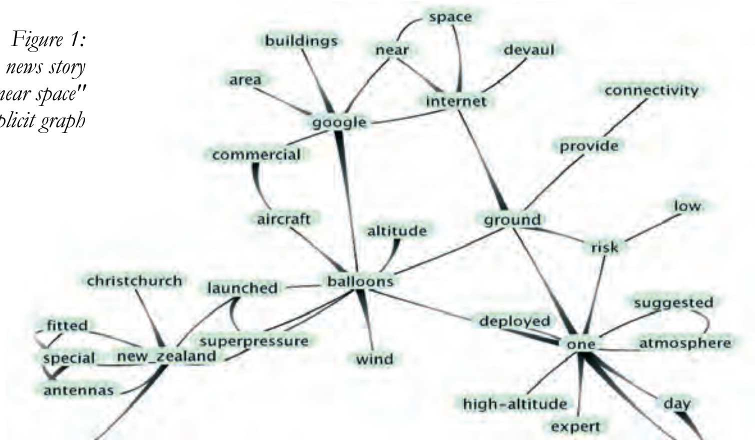
Long-likelihood ratio	Foreground method
plastic - made	superpressure - balloons
months - airborne	launched - new_zealand
suggested - atmosphere	google - balloons
special - fitted	suggested - atmosphere
force - air	force - air

We now propose a new type of word clouds, *diverging (association) clouds,* that aim at helping users to explore the novel associative knowledge emerging from textual documents. Given a search term, the diverging cloud of a document highlights those words that have a special association with the search term. As a motivating application, consider word clouds as summaries of news stories. If the user has a special interest, say 'iPhone', we would first of all like the word clouds to be focused or conditioned on this search term, i.e., only show terms to which 'iPhone' is associated in the news story. Secondly, we would like to see only novel information about the iPhone, not the obvious ones such as 'Apple' and 'mobile'. The diverging clouds aim to do exactly this, directly based on the foreground associations of a news story as a representation of potentially new semantic associations. For a sample of diverging (association) clouds, see Figure 4 in the Results section.

In news, it is very common that the information on a certain event comes in over time. This is even more so for news published on the web or discussed in internet forums. For instance, considering an incident (e.g., the Boston Marathon bombing) which has a large impact and is related to many people, information and updates concerning the event are usually published frequently on news websites. For each news story update, most parts remain the same, some of the information changes, something is added, and something is removed. To get an overall picture of the event, one should go through many articles and collect the new bits of information from each of them while at the same time most of the information is redundant.

Automatic document summarization is a method for overcoming this problem. The main goal of document summarization is to represent the information in a set of documents in a short and possibly non-redundant manner.

Figure 1:
A subset of the foreground associations for the news story "Google tests balloons to beam internet from near space" represented as an explicit graph



Two different approaches have been used in document summarization - generating text using the documents as reference (Hori et al., 2003; Knight et al., 2002) and selecting a representative set of sentences from the documents, e.g., by using Support Vector Machines (Yeh et al., 2005), Hidden Markov Models (Conroy and O'leary, 2001), and Conditional Random Fields (Sheh et al., 2007).

In this paper, we sketch a tentative method which uses sentence extraction for document summarization using foreground associations as reference. In the future, we plan to enhance this method and also combine it with text generation into a hybrid method. Sentence selection is based on the principle that our goal is to cover as much of the foreground associations as possible with the minimum number of sentences. The intuitive idea is that the foreground associations describe the most relevant aspects of the document set. Now, if these associations are covered by a certain set of documents, it is reasonable to assume that we have also captured the essentials of these documents.

A greedy algorithm for selecting the sentences is the following:

1. Select the strongest association *e* from the foreground associations.
2. Let *S(e)* be the set of all sentences which contain both words of *e*.
3. For each sentence *s* in *S(e)* calculate the score of the sentence as the sum of the foreground association weights for all the word pairs found in the sentence.
4. Output the highest scoring sentence *s**.
5. Remove *e* and all other pairs which appear in *s** from the foreground associations.
6. If the size of the summary is not sufficient and the foreground is not empty, go to (1).

The design of the algorithm follows the following two principles. First, the highest weighted pair should contain the most important information. This is the reason why we start looking for the sentences where the words of the highest association in the foreground appear. The second principle is that we do not want to include information which is already included. This is the reason for the step (5) above. If we are interested in penalizing sentences which contain pairs which are already covered, then in step (5) it is possible to change the values of the association weights into negative constants. For our experiments with summary generation on the topic of Google balloons, see the Results section 3.2.2.

3. Results

3.1 Topic analysis

We have applied topic mining to both the English and the Finnish languages. Our first experiment was to construct a model using 2 million articles from the English Wikipedia (at the time this comprised 13 % of the entire English Wikipedia).

The vocabulary is created by extracting eight features from the raw text divided into subdocuments. These features are based on part-of-speech (POS) classification (extracting and lemmatizing nouns, verbs, adjectives and adverbs) and named entity recognition (NER) (tagging words and groups of words as persons, locations, organizations and miscellaneous). For POS tagging we use FreeLing (Padró et al., 2010) and for NER tagging the Illinois Named Entity Tagger (Ratinov and Roth, 2009). The documents and features are forwarded to the model trainer MPCA as a bag-of-words presentation. The MPCA produces the K (in these examples K=50) strongest topics for the user to name. This name is not used for the projection of text against the model (Equation 3), but it associates a numbered topic to a semantically meaningful context, which is essential for humans who exploit the method. Tables 2 and 3 present one of the fifty topics generated. The topic has intuitively been given the name "Space missions". Documents/texts under analysis are projected against the created model in order to find which topics the text is most strongly related to. Feature extraction and bag-of-words presentation are applied to the single document (as is done for the entire corpus in the model creation) before applying Equation 3 to the projection.

Table 2:
The fourteen strongest Named Entity tags for the topic "Space missions". The un-normalized weighting factor corresponds to the incidence of the Named Entity in the particular topic. LOC stands for location, MISC for miscellaneous, ORG for organization, and PER for person. The weight is given at the left side of each word

Weight	LOC	Weight	MISC	Weight	ORG	Weight	PER
14.25	Russia	34.11	Russian	5.71	NASA	1.51	Venus
6.31	Moscow	11.34	Soviet	5.27	Sun	0.77	Ivan
5.01	Earth	6.98	Ukrainian	1.84	Apollo	0.59	Pluto
4.85	Ukraine	2.09	Estonian	1.13	Mars	0.53	Mars
4.28	Soviet Union	1.98	Georgian	1.05	Moon	0.43	Galileo
1.78	Kiev	1.87	Russians	0.96	Saturn	0.42	Mercury
1.70	Estonia	1.56	Latvian	0.75	NGC	0.38	Moon
1.66	Mars	1.05	Soyuz	0.64	Nikon	0.38	Vladimir
1.56	Jupiter	1.03	Belarusian	0.61	ISS	0.35	Ptolemy
1.52	USSR	0.74	Titan	0.57	GPS	0.34	Kepler
1.35	Georgia	0.62	Martian	0.53	ESA	0.31	Lenin
1.30	Belarus	0.62	Gregorian	0.52	Gemini	0.30	Boris
1.31	Latvia	0.54	Chechen	0.50	Canon	0.28	Koenig
1.17	Saint Petersburg	0.50	Earth	0.44	AU	0.28	Star

Table 4 shows an example based on the BBC article entitled "Storm Sandy: Eastern US gets back on its feet" (31 October 2012). Table 4 presents the five strongest topics given by the model for this news article. The numbers in front of the topics are normalized statistical weights of each topic. Table 5 presents the Named Entities given by the NER tagger for this news article.

Table 3: The strongest Part of Speech (POS) tags for the topic "Space missions". JJ stands for adjective, NN for noun, RB for adverb and VB for verb. The weight is given at the left side of each word

Weight	JJ	Weight	NN	Weight	RB	Weight	VB
2.63	solar	2.06	star	15.16	man	2.85	see
1.90	light	1.89	space	2.50	approximately	1.93	take
1.71	lunar	1.37	system	2.34	away	1.52	discover
1.41	html	1.34	planet	1.98	z_times	1.43	show
1.40	russian	1.11	object	1.66	close	1.36	give
1.08	red	1.06	camera	1.56	actually	1.32	move
1.06	astronomical	0.96	light	1.47	relatively	1.30	name
1.05	bright	0.93	satellite	1.46	slightly	1.28	find
1.04	scientific	0.87	crater	1.44	probably	1.26	appear
1.04	black	0.86	day	1.22	roughly	1.13	observe
1.03	dark	0.84	mission	1.20	sometimes	1.11	launch
0.99	similar	0.81	orbit	1.19	currently	1.01	call
0.97	visible	0.80	distance	1.16	long	0.86	refer
0.90	optical	0.75	lens	1.16	directly	0.77	base

Table 4: The strongest topics of the BBC, 31 October 2012 article "Storm Sandy: Eastern US gets back on its feet". The normalization is such that the total sum of the weights of all words in the material is 1

	Weight	Topic Name
1	0.0512	US politics
2	0.0511	Sci-fi and technology
3	0.0391	US traffic and information networks
4	0.0384	Latin America
5	0.0342	Physics

Table 5: The Named Entities for the BBC news article "Storm Sandy: Eastern US gets back on its feet" (31 October 2012)

Freq.	Type	Entity	Freq.	Type	Entity
1	MISC	Democratic	1	PER	Andrew Cuomo
1	MISC	Earth A	1	PER	Barack Obama
1	MISC	Jersey Shore	2	PER	Chris Christie
1	MISC	Nasdaq	2	PER	Christie
3	MISC	Republican	1	PER	Donna
1	LOC	Atlantic City	1	PER	Joseph Lhota
1	LOC	Canada	1	PER	Michael Bloomberg
1	LOC	Caribbean	1	PER	Mitt Romney
1	LOC	Easton	1	PER	Mt Washington
1	LOC	Haiti	2	PER	Obama
1	LOC	Hudson River	1	PER	Paul Adams
1	LOC	JFK	1	PER	Romney
4	LOC	Manhattan	6	PER	Sandy
2	LOC	Maryland	1	ORG	AP
1	LOC	NY City	1	ORG	CNN
1	LOC	New Hampshire	1	ORG	Coriolis Effect
5	LOC	New Jersey	1	ORG	Little Ferry
7	LOC	New York	1	ORG	MTA
2	LOC	New York City	1	ORG	Metropolitan Transit Authority
1	LOC	New York Stock Exchange	1	ORG	Moonachie
1	LOC	New York University	1	ORG	National Weather Service
1	LOC	Ohio	1	ORG	New York Stock Exchange
1	LOC	Queens	1	ORG	Newark Liberty
1	LOC	Teterboro	1	ORG	Tisch Hospital
4	LOC	US	1	ORG	Trams
1	LOC	Washington DC	1	ORG	US Department of Energy

Tables 6 and 7 explore the fifth strongest topic of this news article, "physics", showing a collection of the strongest individual Wikipedia articles on this topic, and strongest features of this topic.

The Finnish language Wikipedia turned out to be far less extensive than the English one. Instead, we used a collection of 73 000 news articles from the Finnish News Agency (STT). Generally, the text in this material is of good quality, but there are some limitations: sports news are dominating and there are very few information technology related news (no Apple, Google, Facebook, Twitter, etc.). The STT news used here date from the years 2002-2005 including also 5000 news from February 2013. For POS tagging the STT news we used a commercial morphological parser, FINTWOL by Ling-Soft Ltd., and for NER tagging we created lists of NER tagged words to which we compared single and groups of POS tagged and lemmatized words. As an example for Finnish, Tables 8 and 9 present results based on an article about the re-election of Giorgio Napolitano as the president of Italy (*Talouselämä*, 22 April 2013).

Table 6: The strongest individual Wikipedia articles for the topic "Physics"

Terahertz time-domain spectroscopy
List of materials analysis methods
Fiber laser
Cryogenic particle detectors
Varistor
Neutron generator
Laser ultrasonics
Optical amplifier
Thyristor
Electric current
Neutron source
Voltage-regulator tube
Switched-mode power supply
Gas-filled tube
Isotopes of plutonium
Superconducting magnet

Table 7: The strongest features for the topic "Physics"

NE-LOC	US, Europe, Chernobyl, Hiroshima, Earth.
NE-MISC	X-ray, Doppler, CO2, °C, CMOS, Fresnel.
NE-ORG	CERN, IPCC, IAEA}.
NE-PER	Maxwell, Edison, Gibbs, Watt, Richter, Einstein, Rutherford, Faraday, Bohr}.
POS-JJ	nuclear, electrical, magnetic, liquid, thermal, atomic, mechanical, solid}.
POS-NN:	energy, power, system, gas, material, temperature, pressure, air, effect, frequency, wave, field, heat, particle, unit, process, signal, mass, device, surface, circuit, light.
POS-RB:	relatively, extremely, slowly, fast.
POS-VB:	produce, require, cause, measure, reduce, increase, generate, allow, apply, create.

Table 8: The five strongest topics for a Talouselämä, 22 April 2013 article (English translation in parenthesis)

Number	Weight	Topic Name
1	0.1014	Vaalit (elections)
2	0.0567	Kansainvälinen konflikti (international conflict)
3	0.0497	Sää (weather)
4	0.0438	Aseellinen selkkaus (armed conflict)
5	0.0434	Tuloneuvottelut (income negotiations)

Table 9: The Named Entities for the Talouselämä, 22 April 2013 article

Freq.	Type	Value
2	MISC	presidentti (president)
1	MISC	radikaali (radical)
2	LOC	Italia (Italy)
1	LOC	maa (country)
2	PER	Napolitano
1	ORG	hallitus (government)
2	ORG	parlamentti (parliament)

Figure 2 shows all the fifty topics obtained for the *Talouselämä*, 22 April 2013 article. The highest peak is the strongest topic "president". The number of Named Entities for the example in Finnish is much smaller than that in English. The state of the art NER taggers for Finnish are not as evolved as the taggers for English.

The overall results are, in fact, better for the BBC article; there are more NER tagged words and the strongest topics correspond better to the semantic contents of the article.

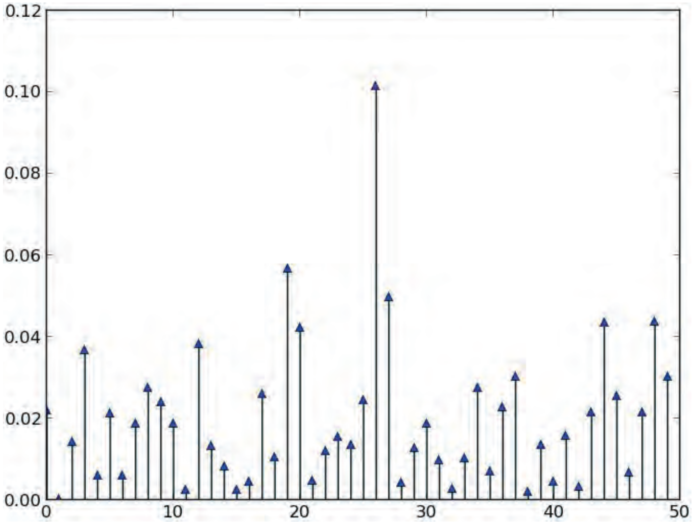


Figure 2: The fifty strongest topics for one news article projected against model created with STT news data. The horizontal axis shows the number of the topic and the vertical axis shows normalized weight of the topic

However, the results using the model created with STT data are far better than those created with the Finnish Wikipedia. This is demonstrated in Figure 3 where the strongest topics do not as strongly rise above the rest and, furthermore, the five strongest topics are mostly not significant: Finnish politics, philosophy and religion, natural sciences, computer games, and banks and monetary policies. This shows that the corpus and named entity data used to create the model is sufficiently extensive and of good quality.

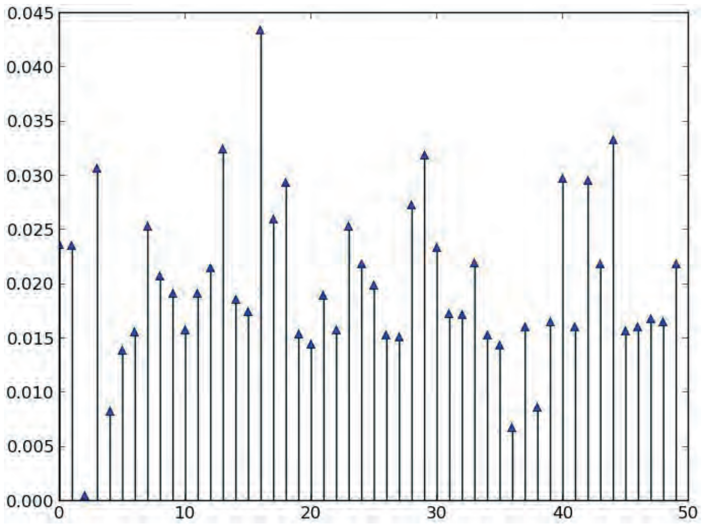


Figure 3: The fifty strongest topics for one news article projected against a model created using the Finnish Wikipedia

As another source for the corpus in Finnish we used the free newspaper sheet *Metro*. For POS tagging of *Metro* news we used the Open Source Morphology for Finnish, OmorFi (Lindén et al., 2011), and for NER tagging we used a combination of OMorFi and our own POS tagging version created for STT news.

3.2 Word association analysis

3.2.1 Diverging word clouds

In this section we present some results of the Software Newsroom applications that use word association analysis as their basis. As before, for single document experiments we use the English Wikipedia as the background corpus and a story from BBC, "Google tests balloons to beam internet from near space" (Kelion, 2013), as the foreground document that we are interested in.

Given a document *d* and a word *w*, a specification for the corresponding diverging association cloud is directly obtained from the foreground associations of the document: take the top *n* words associated with *w* in the foreground and position them in the word cloud according to their weights. Figure 4 illustrates the idea using the document on Google balloons.

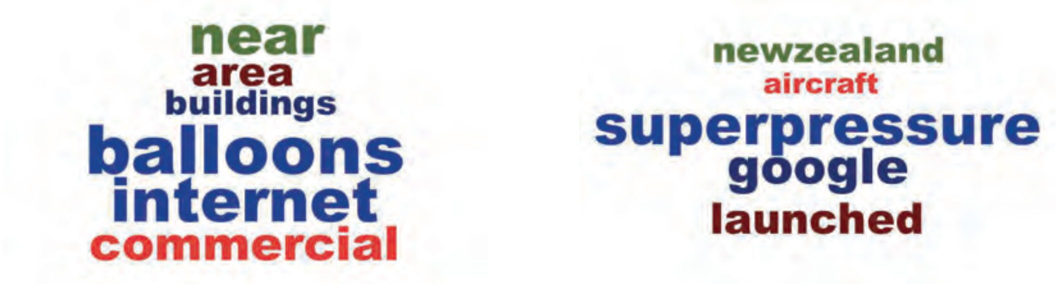


Figure 4: The diverging word cloud created from the foreground associations of the news story "Google tests balloons to beam internet from near space". The search term for the left diverging cloud is 'google' and for the right diverging word cloud the search term is 'balloons'. We used an internet tool (Word Clouds for Kids, 2012) for generating the word clouds

These association clouds give a good idea of what the document could be about. Such word clouds could also be implemented in an interactive manner: as the user clicks on a word in the cloud, the selected word becomes the next search term and, correspondingly, all divergent clouds are re-rendered for all documents using the new focus. A drawback of this method is that it takes time to get used to the fact that the word cloud is conditioned on the search term and thus interpreting the results could be non-intuitive for novice users. In order to alleviate the problem, it would be possible to also present the search term together with the words but currently we do not have a clear idea of how to present this in an intuitive manner.

3.2.2 Summary generation

For our experiments with the summary generation algorithm presented earlier, we collected four news stories on the same topic from different news sources - BBC: "Up, up and away: Google to launch Wi-Fi balloon experiment" (Kelion, 2013), National Geographic: "Google's Loon Project Puts Balloon Technology in Spotlight" (Handwerk, 2013), ARS Technica: "Google's balloon-based wireless networks may not be a crazy idea" (Brodkin, 2013), CNN: "Google tests balloons to beam internet from near space" (Smith-Spark, 2013).

For the background associations we used Wikipedia news stories and the foreground associations were calculated using the respective news stories. We then applied the algorithm which we described on this set of data. The total number of words in all the four documents was 3 696.

The first four sentences, containing a total of 120 words, returned by the algorithm were the following:

- Google is reportedly developing wireless networks for sub-Saharan Africa and Southeast Asia that would combine a technology well established for such purposes (TV White Spaces) with one that's a bit more exotic - balloons that transmit wireless signals. - *ARS Technica*
- Project Loon balloons are made of plastic just 3 mm (0.1in) thick, another Orlando-based firm, World Surveillance Group, sells similar equipment to the US Army and other government agencies. - *BBC*
- It has been working on improving connectivity in the US with Google Fiber and bringing the internet to underserved populations overseas through White Spaces networks. - *ARS Technica*
- A company called Space Data makes balloon-based repeater platforms for the US Air Force that "extend the range of standard-issue military two-way radios from 10 miles to over 400 miles." - *ARS Technica*

The application of the algorithm yields promising results. Our next goals are improving and evaluating the current method. It is important to note here that the way the extracted sentences are presented to the user is also a very important aspect. For instance, consider the third sentence which has a co-reference resolution problem (i.e., the sentence starts with "it" and we do not know what "it" is). In such cases it makes sense to present consecutive sentences together in the summary regardless how they are ordered by the algorithm. In some cases this could help to overcome the co-reference resolution problem. It is also possible to provide some context to the user, for instance, when the user's cursor hovers over an extracted sentence, the sentences which are before and after it in the news story can be shown.

4. Discussion

Application of MPCA seems to work well for news search by topic analysis. It is likely that also other variants of probabilistic modeling perform well for news identification. Our second approach, association analysis, also clearly enhances the effectiveness of the "news nose". A question then arises, whether other methods could be effective as well, or even better than the adopted approaches.

In contrast to statistical methods such as PCA, *cluster analysis* can best be seen as a heuristic method for exploring the diversity in a data set by means of pattern generation (van Ooyen, 2001). Cluster analysis may be applied for finding similarities and trends in data (described using the common term *pattern recognition*). An example of cluster analysis is the *expectation maximization* (EM) algorithm, which has recently been applied to astronomical data for identifying stellar clusters from large collections of infrared survey data (Solin, Ukkonen and Haikala, 2012). Cluster analysis has also been used in, e.g., market research within a more general family of methodologies called *segmentation methods*. These can be used to identify groups with common attitudes, media habits, lifestyle, etc. Cluster analysis is probably less well suited for news search than probabilistic models like MPCA, since the semantic contents of articles that contain more than one topic are not resolved by cluster analysis (e.g., Newman at al., 2006), while probabilistic modeling clearly performs well in such cases provided that the corpus and named entity data used for the model creation are sufficiently extensive. This will result in only a small number of unrecognized words that cannot be tagged, and thus a high resolving power of topics and named entities.

Supervised learning methods divide objects such as text documents into predefined classes (Yang, 1999). Cluster analysis and PCA are data driven methods which can extract information from documents without *a priori* knowledge of what the documents may contain (Newman et al., 2006), and topic categorization (i.e., a topic model) is created by the algorithm without rules or restrictions on the contents of a topic, which is why such methods are called *unsupervised learning*. Obviously supervised learning is poorly suited for news search from arbitrary textual data, since the topics of potential news in the material cannot be predicted, and it is thus impossible to recognize new emerging topics.

A further, more advanced analysis of complex data may incorporate the use of *semantic networks*. Methods of this category are *Traditional* and *Improved Three-Phase Dependency Analysis* (TTPDA, ITPDA). These algorithms have been applied to recognition of semantic information in visual content and they use Bayesian networks to automatically discover the relationship networks among the concepts. These methods can be applied, for example, to automatic video annotation. (Wang, Xu and Liu, 2009).

In this paper, we have mainly interpreted the associations on a single association level rather than as a network. But these associations, both background and foreground, can also be seen as a kind of semantic network where words are nodes and the edges represent the associations. Analyzing the background associations as a network might give interesting results in automatic word domain discovery or for finding interesting subnetworks that connect two words. The same applies for the foreground associations, which might provide interesting inference and application possibilities when interpreted as word networks and used as such. Thus, in the future, our models and methods could be improved in their accuracy. More efficient, scalable algorithms could be designed and, perhaps more interestingly, additional novel applications could be invented with help of the background and foreground models, especially in the broad areas of information browsing and retrieval.

Considering the topic model and data used for the training, our experiments indicate that the comprehensiveness, quality, and also the semantic similarity of the text corpus and named entity data with the data under analysis are critical to the effectiveness of the search algorithm. This is of course obvious, but poses a challenge for automated news search since language evolves and the language used in, e.g., social media that obeys no standard rules diffuses with an increasing speed to various media channels. Should we accept this and modify the models and additionally also adopt slang in the presentation of news, or try to force the users to educate themselves in order to write in decent standard language also in social media?

An aspect of crucial importance in (automated) news search is the quality of the data. The internet is full of hoaxes and distorted information, and finding assurance for the reliability of potential news may sometimes be challenging, and will require too much time.

This may lead to that the potential news becomes yesterday's news or that it is published by a competitor before sufficient background information is found. The Software Newsroom should therefore trace all possible metadata on the sources, time, places, people, and organizations associated with the creation of the information found by automated means. While this cannot rely merely on software, automation can be used to significantly improve the effectiveness and speed of the process.

5. Conclusions

We have developed and applied methods for automated identification of potential news from textual data for use in an automated news search system called Software Newsroom. The purpose of the tools is to analyze data collected from the internet and to identify information that has a high probability of containing news. The identified potential news information is summarized in order to help understanding the semantic contents of the data and also to help in the news editing process.

Two different methodological approaches have been applied to the news search. One method is based on topic analysis which uses MPCA for topic model creation and data profiling. The second method is based on association analysis that applies LLR. The two methods are used in parallel to enhance the news recognition capability of Software Newsroom.

For the topic mining we have created English and Finnish language corpora from Wikipedia and several Finnish language corpora from Finnish news archives, and we have used bag-of-words presentations of these corpora as training data for the topic model. We have made experiments of topic analysis using both the training data itself and arbitrary text parsed from internet sources. The selected algorithmic approach is found to be well suited for the task, but the effectiveness and success of news search depends strongly on the extensiveness and quality of the training data used for the creation of the topic model. Also, semantic similarity of the target text with the corpus used for the model creation generally improves the search effectiveness. The large difference between the language commonly used in user-created internet content and standard language poses a challenge for news search from social networks, since a significant part of the language is not recognized by the part-of-speech and name entity taggers. A simple solution for this would be a translator that would preprocess the unknown slang words, turning them into standard language. Another would be a slang-based corpus. The latter has the disadvantage that the resulting raw news material would be composed of slang and it would have to be translated into standard language before publishing. Thus, our plan is to collect a small dictionary of the most common words used in social media and use them for further experiments on social media.

In the association analysis we have used a methodology for detecting novel word associations from a set of documents. For detecting novel associations we first used the background corpus from which we extracted such word associations that are common. We then collected the statistics of word co-occurrences from the set of documents that we are interested in, looking for such associations which are more likely to appear in these documents than in the background.

We also demonstrated applications of Software Newsroom based on association analysis - association visualization as a graph, diverging (association) clouds which are word clouds conditioned on a search term, and a simple algorithm for text summarization by sentence extraction. We believe that the background-foreground model has significant potential in news search. The simplicity of the model makes it easy to implement and use. At the same time, our experiments indicate great promise in employing the background-foreground word associations for different applications.

The combination of the two methods has not yet been implemented. This is in our plans for the near future. and the application of both methods on social media using a pre-translator of social media language is underway. Potential future work also includes experiments on automated news generation and application of our methods for other purposes, e.g., improvement of recommendations.

Acknowledgements

We are grateful to Jukka Mauno from MTV3 for raising the idea of Software Newsroom. We acknowledge Tekes, the Finnish Funding Agency for Technology and Innovation, and Tivit Oy for financing this research. Especially we wish to thank the Focus Area Director of the Next Media research programme, Eskoensio Pipatti, for his support. This work has also been supported by the Algorithmic Data Analysis (Algodan) Centre of Excellence of the Academy of Finland and Hecse, Helsinki Doctoral Programme in Computer Science - Advanced Computing and Intelligent Systems. Finally, we thank the referees for valuable comments.

References

Agrawal, R., Imieliński, T. and Swami, A., 1993. Mining association rules between sets of items in large databases. *ACM SIGMOD Record.* 22(2), pp. 207-216

Allan, J., 2002. *Topic Detection and Tracking: Event-based Information Organization*. Dordrecht: Kluwer Academic Publishers

Berry, M.W., Dumais, S.T. and O'Brien, G.W., 1994. Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review,* 37(4), pp. 573-595

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3, pp. 993-1022

Brodkin J. (2013). Google's balloon-based wireless networks may not be a crazy idea. *ARS Technica*, June 2, 2013. [Online] Available at: <http://arstechnica.com/information-technology/2013/06/googles-balloon-based-wireless-networks-may-not-be-a-crazy-idea/>. [Accessed 26 June 2013]

Buntine, W. and Jakulin, A., 2004. Applying discrete PCA in data analysis. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (UAI2004), pp. 59-66

Buntine, W. and Jakulin, A., 2006. Discrete component analysis. *Subspace, Latent Structure and Feature Selection, Lecture Notes in Computer Science*. Vol. 3940, pp. 1-33

Church, K.W. and Hanks, P., 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1), pp. 22-29

Comintelli, 2013. [Online] Available at: <http://www.comintelli.com/Company/Press-Releases/Esmerk-launches-new-current-awareness-platform-Esm> [Accessed 31 October 2013]

Conroy, J.M. and O'leary, D.P. 2001. Text summarization via hidden markov models. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval,* pp. 406-407

Dumais, S. T., Furnas, G.W., Landauer, T.K., Deerwester, S. and Harshman, R., 1988. Using latent semantic analysis to improve access to textual information. *Proceedings of the SIGCHI conference on Human factors in computing systems,* pp. 281-285

Dunning, T., 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1), pp. 61-74

Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T. R., and Alho, I., 2004. *Iso suomen kielioppi.* Helsinki: Suomalaisen Kirjallisuuden Seura (Available online at http://scripta.kotus.fi/visk/etusivu.php)

Handwerk, B. (2013). Google's Loon Project Puts Balloon Technology in Spotlight. *National Geographic*, June 18, 2013. [Online] Available at: <http://news.nationalgeographic.com/news/2013/06/130618-google-balloon-wireless-communication-internet-hap-satellite-stratosphere-loon-project/>. [Accessed 26 June 2013]

Harris, Z., 1954. Distributional Structure. *Word*, 10(23), pp 146-162

Hofmann, T., 1999. Probabilistic Latent Semantic Indexing, *Proc. 22nd Annual International SGIR Conference on Research and Development in Information Retrieval*, pp. 50-57

Hori, C. and Furui, S., 2003. A new approach to automatic speech summarization. *IEEE Transactions on Multimedia*, 5(3), pp. 368-378

Huang, S., Peng, X., Niu, Z. and Wang, K., 2011. News topic detection based on hierarchical clustering and named entity. *7th International Conference on Natural Language Processing and Knowledge Engineering.* pp. 280-284

Hynönen, T., Mahler, S. and Toivonen, H., 2012. Discovery of novel term associations in a document collection. *Bisociative Knowledge Discovery, Lecture Notes in Computer Science,* Vol. 7250, pp. 91-103

Karlsson, F., 1990. Constraint grammar as a framework for parsing running text. *Proceedings of the 13th Conference on Computational Linguistics*, Vol. 3., pp. 168-173

Kelion L. (2013). Google tests balloons to beam internet from near space. *BBC*, June 15, 2013. [Online] Available: <http://www.bbc.co.uk/news/technology-22905199>. [Accessed 26 June 2013]

Kimura, M., Saito, K. and Uera, N., 2005. Multinomial PCA for extracting major latent topics from document streams, *Proc. 2005 IEEE International Joint Conference on Neural Networks*, Vol. 1, pp. 238-243

Knight, K. and Marcu, D., 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1), pp. 91-107

Lindén, K., Axelson, E., Hardwick, S., Pirinen, T.A. and Silfverberg, M., 2011. HFST-Framework for Compiling and Applying Morphologies. In: Mahlow, C. and Piotrowski, M., eds.(2011). *Systems and Frameworks for Computational Morphology. Communications in Computer and Information Science*, Vol. 100. Berlin-Heidelberg: Springer. pp. 67-85

Lindén, K. and Pirinen, T., 2009. Weighted finite-state morphological analysis of Finnish compounds. In: Jokinen, K. and Bick, E., eds.(2009). *Proc. Nordic Conference of Computational Lingustics.* Odense: NEALT

McDonald, D. D., 1996. Internal and external evidence in the identification and semantic categorization of proper names. In: Boguraev, B. and Pustejovsky, J., eds. (1996). *Corpus Processing for Lexical Acquisition.* Cambridge, MA: MIT Press. pp. 21-39

Meltwater, 2013. [Online] Available at: <http://www.meltwater.com/products/meltwater-buzz-social-media-marketing-software/> [Accessed 31 October 2013]

Miller G.A., 1995. WordNet: A Lexical Database for English, *Communications of the ACM,* 38(11), pp. 39-41

Nadeau, D. and Sekine, S., 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1), pp. 3-26

Newman, D., Chemudugunta, C., Smyth, P. and Steyvers, M., 2006. Analyzing Entities and Topics in News Articles using Statistical Topic Models. *Lecture Notes in Computer Science*, Volume 3975, pp. 93-104

van Ooyen, A., 2001. Theoretical aspects of pattern analysis. In: L. Dijkshoorn, K. J. Tower, and M. Struelens, eds. *New Approaches for the Generation and Analysis of Microbial Fingerprints.* Amsterdam: Elsevier, pp. 31-45

Padró, L., Reese, S., Agirre, E. and Soroa, A., 2010. Semantic Services in FreeLing 2.1: WordNet and UKB. *Proceedings of the Global Wordnet Conference 2010*

Pearson, K., 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(11), pp. 559-572

Ratinov, L. and Roth, D., 2009. Design Challenges and Misconceptions in Named Entity Recognition. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pp. 147-155

Salton, G., 1991. Developments in Automatic Text Retrieval. *Science*, Vol 253, pp. 974-979

Shen, D., Sun, J.T., Li, H., Yang, Q. and Chen, Z., 2007. Document summarization using conditional random fields. *Proceedings of the 20th international joint conference on Artifical intelligence,* Vol. 7, pp. 2862-2867

Smith-Spark L. (2013). Up, up and away: Google to launch Wi-Fi balloon experiment. *CNN*, June 15, 2013. [Online] Available at: <http://www.bbc.co.uk/news/technology-22905199>. [Accessed 26 June 2013]

Solin, O., Ukkonen, E., and Haikala, L., 2012. Mining the UKIDSS Galactic Plane Survey: star formation and embedded clusters, *Astronomy & Astrophysics*, Volume 542, A3, 23 p

Toivonen H., Gross, O., Toivanen J.M. and Valitutti A., 2012. Lexical Creativity from Word Associations. *Synergies of Soft Computing and Statistics for Intelligent Data Analysis. Advances in Intelligent Systems and Computing.* Vol. 190, pp. 17-24

Wang, F., Xu, D. and Liu, J., 2009. Constructing semantic network based on Bayesian Network. *1st IEEE Symposium on Web Society*, pp. 51-54

Word Clouds for Kids, 2013. [Online] Available at: <http://www.abcya.com/word_clouds.htm>. [Accessed 26 June 2013]

Yang, Y., 1999. An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, Vol 1, pp. 67-88

Yeh, J.Y., Ke, H.R., Yang, W.P. and Meng, I., 2005. Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing & Management*, 41(1), pp. 75-95

JPMTR 023 | 1316
UDC 659.3:81'37

Research paper
Received: 2013-08-21
Accepted: 2013-10-30

Portable profiles and recommendation based media services: will users embrace them?

Asta Bäck and Sari Vainikainen

VTT Technical Research Centre of Finland
Vuorimiehentie 3
P. O. Box 1000
FIN-02044 VTT, Finland

E-mails: asta.back@vtt.fi
sari.vainikainen@vtt.fi

Abstract

User data and user profiles are very important in current web based business models and applications. Advertising revenue is to a large extent generated based on implicit and explicit user data. We propose the use of semantic, portable and user-controllable profiles to capture and model user data that can be used for personalising media services and particularly for making recommendations.

In this paper, we present results from four user tests where users created semantic profiles and received personalised recommendations based on these profiles. We have studied users' expectations and requirements for profile portability as well as how users have experienced the creation of a semantic profile using different data sources. We have also studied how users experienced recommendation based media services in connection with user-controlled interest profiles. In all four test cases, users have created the profiles and received the recommendations using the prototypes we have developped. User feedback was gathered through interviews and web surveys.

Users welcomed the idea of being able to control their profile data but they also had questions and concerns about privacy if the profile is shared between services. Users had dual concerns about the recommendations: some users were afraid of a too limited view if the service only relies on the user's profile; some users were afraid of being overwhelmed by too many recommendations.

Keywords: user profiles, portable profiles, media services, recommendations, semantic web technologies, linked data

1. Introduction

1.1 Background

User data has become the driving force of web based services, where the knowledge of user behaviour and interests is being turned into revenues by linking advertisers to potential customers. Google and Facebook with their free end user services and refined tools for advertisers are the prime examples of this model. From the end user point of view, free services still come with a price: every user activity is logged and analysed and the user has little influence on and knowledge of what data is stored about her. From the business perspective, the big players get a huge advantage over smaller ones, as they can accumulate versatile data about consumers and use this information to generate advertising revenue.

An alternative approach to user data has been proposed by several initiatives, such as the VRM project at Harvard (Project VRM, n.d.), and the World Economic Forum (World Economic Forum, 2013) as well as the British midata initiative (GOV.UK, 2013).

The core idea of these initiatives is that the business model with personal data is turned upside down: the users are made into owners of their own data and given the opportunity to offer this data to various service providers if and when they see value in sharing their data.

Users are not the only ones to benefit from this: rich portable profiles would give smaller players the opportunity to personalise their offering to users and in this way make their services more attractive and relevant to users.

The vision of user controlled personal data is both promising and challenging. First of all, the potential amount of personal data is almost unlimited, as can be concluded from the following categorisation (World Economic Forum, 2011):

1. Personal Attribute Data:
   Data about the attributes of an individual;
2. Volunteered Data:
   Data created and explicitly shared by individuals. (e.g., social network profiles);
3. Observed Data:
   Data captured by recording actions of the individuals, such as click streams and transactions or location data when using cell phones;
4. Inferred Data:
   Data about individuals, based on the analysis of personal, volunteered and/or observed information.

Many questions arise. Would people be willing to manage their data and share it with various service providers? Are companies ready to change the way they acquire and manage user data? In this new model, the main competitive edge of a company would be the skills and methods to utilise the data that the user offers, not the ownership of the data.

We have approached user controllable personal data using the concept of portable semantic user-controlled interest profiles, and we have implemented a Semantic Portable Profile Platform (SP3) for this purpose. Using semantic web technologies and linked data it is possible to enrich and share profile information across services so that the services will understand the profiles in a similar manner. Linked data can also be used to enrich the user profile information with additional relevant information which means that small amounts of data can be used for making relevant personalisation and recommendations.

Others have also started to explore these opportunities. Kay and Kummerfeld (2013) have proposed the scrutinability of one's user model as a way to increase the quality of personalisation and user experienced trust towards a recommendation and personalisation system. They particularly look at long-term user models that accumulate during months and years of using a particular application, but they do not address the issue of portability.

Bojars et al. (2008) have addressed the technical aspects of transferring user data between social networking applications and proposed a semantic web technologies based solution utilising the FOAF (Friend-of-A-Friend)[1] and SIOC (Semantically Interlinked Online Communities)[2] vocabularies. Orlandi, Breslin and Passant (2012) propose a methodology for automatic creation and aggregation of interoperable and multi-domain user profiles of interest using semantic technologies such as linked data. They combine expressions of interests from multiple social media sites and rank the interests based on their frequencies in different data sources in order to make inferences concerning the relative importance of different interests.

In our method, we also utilise data from social media services in creating the profile but our focus is on the combination of creating a profile and utilising it in media recommendations. Heitmann et al. (2010) address the problem of preserving user privacy while integrating multiple information sources and present an architecture for privacy-enabled user profile portability based on existing standards with a use case from the e-Health domain.

Portable semantic interest profiles would make it easy for media companies to offer various kinds of personalised services, most of all personalised recommendations.

Recommendations are seen as an increasingly important part of many web based services. They are mainly used to reduce the effort required from the user to find items that are most likely to interest them and to reduce the information overload (Liang, Lai and Ku, 2006).

Recommendations are often studied by looking at the precision of the recommendation algorithms but, in real life applications, many other aspects play a role in how the users will experience the recommendations and what their impact is on the whole usage experience of a recommendation service (Knijnenburg et al., 2012; Konstantin and Riedl, 2012).

We have created a number of applications that use semantic user profiles for generating recommendations in order to study various aspects of the topic. In this paper, we will present our results relating to profile portability from the end users' point of view: what are users' expectations and requirements for profile portability and how do they experience the creation of a semantic profile using different data sources.

Our main use case for the portable profiles has been recommendations. Recommendations are an important application area for profiles and they give the test users a practical example of how the profiles can be utilised. In this paper, we report on how users have experienced recommendation based media services in connection with user-controlled interest profiles.

We first present our framework for semantic portable profiles. We then present the cases where semantic profiles have been tested with working prototypes and test users.

We will also present our results relating to profile creation - what kind of profiles users created in the different cases and their experiences of profile creation, acceptance of the concept of portable profile, and views on recommendation based services.

2. Framework and methods

2.1 The semantic portable profile platform

Our Semantic Portable Profile Platform (SP3)[3] supports creating, managing and utilising semantic portable user profiles (Figure 1). The platform consists of tools and methods that allow users to create a semantic interest profile either manually or by importing their data from external social media services.

The platform includes methods for semantic enrichment of metadata, and for linking semantic interests, context and content metadata to make recommendations.

The word portability refers to the opportunity of third parties to utilize the user profile if the user gives his or her acceptance using OAuth (Open Standard for Web Authorisation)[4].

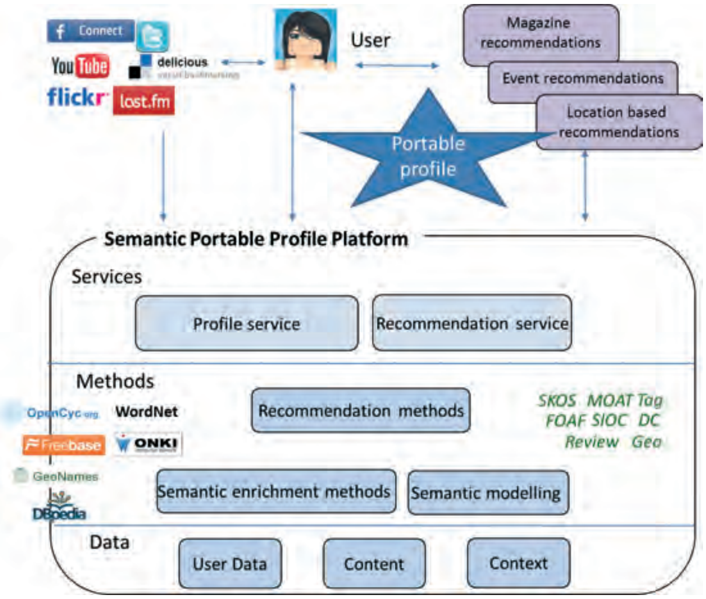


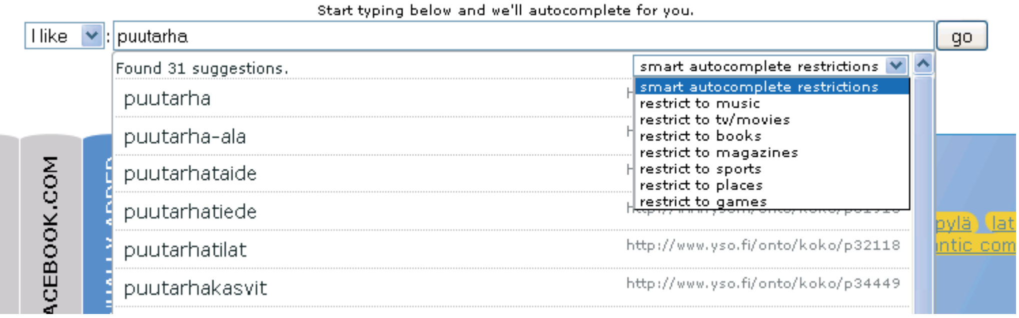
Figure 1: Semantic Portable Profile Platform (SP3)

The profile data is available via REST API[5] in JSON (JavaScript Object Notation)[6] or XML (Extensible Markup Language)[7] format. The user data is also available in RDF (Resource Description Framework)[8].

The user model has been defined reusing concepts and properties from the following ontologies: FOAF, vcard[9], geo[10], Review[11], tags[12], SCOT[13] and DC (Dublin Core), consisting of Dublin Core Metadata Element Set[14] and DCMI Metadata Terms[15]. The content model additionally uses MOAT (Meaning of a Tag Ontology)[16] and SKOS (Simple Knowledge Organization System)[17].

A semantic interest profile consists of items that are defined as a label with text and a URI (Universal Resource Identifier)[18] that is linked to a concept in a linked data dataset.

The user can link his/her interests manually to their semantic meanings using the available tagging widget: after the user has typed the first three characters of a concept name, suggestions will be fetched from selected ontology vocabularies and shown to the user (Figure 2). When the user selects one of the suggested meanings, it will be included in the profile as a Linked Open Data URI.


Start typing below and we'll autocomplete for you.

I like : puutarha go

Found 31 suggestions.

smart autocomplete restrictions
- smart autocomplete restrictions
- restrict to music
- restrict to tv/movies
- restrict to books
- restrict to magazines
- restrict to sports
- restrict to places
- restrict to games

puutarha
puutarha-ala
puutarhataide
puutarhatiede
puutarhatilat http://www.yso.fi/onto/koko/p32118
puutarhakasvit http://www.yso.fi/onto/koko/p34449

Figure 2: Semantic tagging widget for profile creation with automatic suggestions from semantic databases

The widget can be configured to use one or several ontologies or their subsets in any combination. The profile may also include items that are not linked to a semantic meaning. It is also possible to import data from a user's accounts in social media services, such as Delicious, Flickr, YouTube, Last.fm, Facebook and Twitter, to reduce the manual effort of profile creation. The data may consist of demographic information, tags or lists of favourite items, depending on the service.

The platform analyses the imported data and finds the most likely semantic meaning of the found concepts. The results are shown to the user who can correct the meaning, if needed. The user can also modify the results in the same way as manually created interests: indicate the strength of the interest (love, like, hate), or to remove them from the profile.

2.2 Test cases

We have tested semantic profile creation and the concept of portable profiles as well as recommendation based services in the four main cases: magazine recommendations, event recommendations (two separate cases), and videos and TV programmes emphasising location information in making recommendations. A summary of the tests performed is presented in Table 1.

The magazine recommendations case was investigated online. Test users were invited by email: 337 persons created a profile for themselves and 119 of them completed the entire user test by filling out the final questionnaire. Users created their profiles using our semantic widget where they could either select the suggested semantic concepts or write in any words to describe their interest. The profile creation form contained eleven category fields encouraging users to report many different matters, from interests to dreams and problems. Some examples were given to help users understand what kind of input was expected. The option of linking to social media services such as Delicious and YouTube was available but it was not used by many. After completing the profile, users were shown a list of recommended magazine articles. The users were asked to look at the recommended articles and rate how well they matched their interests.

Event recommendations case I was a study carried out as a laboratory test with 15 test users. As a first step, the concept of the service independent profile and the event recommendation service were explained to the users and they were given ten minutes to familiarize themselves with the actual event service. After this, each test user created an interest profile using our semantic tagging widget which included four fields: interests, favourite music, favourite movies and favourite books. The users could enter either free text or the semantic concepts suggested on the basis of the Finnish KOKO ontology of the Finnish Ontology Library Service ONKI[19] as well as the Freebase[20] and Geonames[21] databases. It was possible to import tags from Delicious, YouTube, Flickr and Last.fm but, like in the magazine case, only few test users did so. Each user received 20 recommendations based on their profiles and were asked to rate the relevance of the recommendations. Finally, they filled out a separate feedback questionnaire about the service and profile concepts and profile creation.

Event recommendation Case II was based on a prototype consisting of a mobile phone and an HbbTV (Hybrid Broadcast Broadband TV) event application for showing recommendations. The profiles were created on the profile service web page. Both recommendation applications used the same user profile and the EvenemaX[22] event database containing 1 500 events. In addition to the interests registered in the profile, also the locations of the events and of the user were taken into consideration in making recommendations. Users could rate the recommendations and event information and indicate which events they were planning to attend. They could also see which events their friends were going to attend. We had four test users.

In the *location based recommendation* case, an iPhone application, called Mediatutka (Media radar, in English), was developed and tested. As the first step in the Mediatutka development process we gathered input from potential users in the Owela[23] online co-creation environment (Friedrich, 2013). Three sections were available: one with seven predefined discussion topics, one with four user stories and one for open ideation.

During the three weeks long co-creation period, end users proposed 17 new ideas or started new discussion topics. In total, 426 comments were written in the discussions. 74 users (64 end users, 6 developers and 4 researchers) participated in ideation and discussion. These Owela participants were different from those who tested the actual Mediatutka application.

The implemented Mediatutka app generates personalised recommendations and notifications out of the available content pool. In our tests, the content consisted of TV program data from skimm.tv[24], Stadi.TV's[25] local, citizen-made video clips, and information of new videos from the HelMet)[26] library service. Only Stadi.TV videos could be watched immediately on the phone; skimm.tv recommendations dealt with TV programmes to be broadcast in near future. HelMet videos are in physical format and they need to be fetched from a library; an alert was sent to the user when he or she was closer than 2 km from a public library from where a video matching his or her interests could be borrowed.

As the first step when starting to use the Mediatutka application, users log in with their Facebook credentials and their profile data is imported from their Facebook accounts.

After that, they can add more interests manually. Mediatutka gives recommendations to the user with a single click of a button and also sends location-sensitive alerts in near-real-time. Recommendations and notifications are generated based on user created semantic profiles, user-made ratings of the earlier recommended content and the user's current location. There were 17 test users of the application.

Table 1: The cases where recommendation based media services and the concept of portable profile were tested

Case name	Case description	Profile creation	Testing
Magazine article recommendations	Recommendations were made out of 606 magazine articles originally from 12 magazines from Sanoma Magazines based on user created profiles and manually created content metadata.	Online test. Users created the profiles on their own either by using any words or by choosing words that were linked to a semantic database after typing initial letters. Importing data from Delicious and YouTube was offered but used only by very few. Users were encouraged to report many different things like in a friendship book, from interests and dreams to problems.	337 users created a profile; 119 of them completed the whole user study. Recruited by email.
Event recommendations I	Events from EvenemaX event database. Recommended out of 1 500 entries based on user created profiles and text index of event descriptions.	Laboratory test. Users created their profile online and received as a result 20 event recommendations. Users were asked to express interests as well as favourite music, films and books. These could be expressed as any words or by selecting suggested items after typing initial letters.	15 users.
Event recommendations II	Event recommendations from EvenemaX event database.	Profiles were created on a separate web page (any words + semantic suggestions based on the first typed letters). Facebook import available. Four weeks' test period with events from an event database. Focus group sessions before and after the test period. Owela feedback during the test period. Also an HbbTV event recommendation application was laboratory tested.	Four test users.
Mobile profile creation and location based recommendations and notifications	Mediatutka, an app that gives personalised recommendations and location specific notifications out of 1 200 Stadi.TV videos, 140 HelMet videos and 3 300 TV programmes from skimm.tv during the test period. All steps from profile creation and viewing recommendations using one mobile app.	Profile data is imported from Facebook and additional interests may be added manually and linked to semantic concepts. 167 Facebook interests were analysed; on an average 7 interests per user. Half of the test users (13/26) added interests also manually (100 interests, or 7.7 interests/user).	Initial concept and use case co-development in the Owela online environment (64 users). Mediatutka application testing: 26 users installed the application and logged in at least once. 17 provided feedback after the test period of two weeks.

The majority of our test users have been young adults, with a female majority. In the magazine article recommendation case, the age distribution was fairly wide from 18 to over 64, but the majority was fairly young with 58% of the participants under 35 years of age. Three of the 119 test persons were male. In the event recommendation case I, all 15 test users were between 18 and 23 years old; ten women and five men. In the event recommendation case II, all four test users were young men in their early twenties. In the Mediatutka application user test, 65% of the test users were under 35 years of age; 12 women and five men.

Out of 64 the Owela co-development phase participants, 47% were women and 53% men. Their age range was from 20 to 74 with only 23% being under 39 years of age.

3. Results

3.1 Users profiles and profile creation

In the magazine article recommendation case, 337 user profiles with 4 892 interests were created (14.5 interests/user). Almost two thirds (64%) of the added interests were linked to an ontology, but there were also free text interests in the profiles without semantic linking even though such a link would clearly have been available. 'I'm interested in', 'I spend a lot of time with' and 'Important to me', were the categories of most entries. Each of the eleven categories received at least 200 entries. Some users had not realised that a comma was needed to separate interests from each other and had produced by mistake very long words with many interests combined. Some interests had been impossible to describe with only one tag; e.g. child's problems at school; more time for myself. This type of interests came up in categories concerned with dreams or problems in the user's life.

User assessments of the profile creation were given on a five step scale from +2 (very positive) to -2 (very negative) with the following results:

- 55 % of the users experienced the profile creation as very or fairly easy,
- 65 % indicated that finding the correct term among the suggested terms worked very or fairly well,
- 59 % considered the semantic tag suggestions very or fairly useful.

In the free text comments, some users complained about having been offered several concepts with the same title and not knowing which one to use. Free text comments also revealed some problems and misunderstandings related to the automatic suggestions: one user had understood that the terms represented the available content, which was not the case as the available semantic databases included a huge number of concepts - much more than what was available in the magazine articles of the test.

In the event recommendation case I, only four fields were available in the profile creation widget. Regardless of that, the number of expressed interests was practically the same in both cases: 15.7 interests/user in the event recommendations case I and 14.5 in the magazine case.

The event profile creation form guided users to give more specific information, since they were asked to tell about their favourite music, movies and books. Users did mention specific artists and items but, regardless of that, the most popular tags were comedy, dance, music, and pop music. Many identical interests were mentioned in both cases, as can be seen in Table 2.

In the first tests, only very few participants utilised the opportunity to bring in data from their social media accounts. A big obstacle was that very few had accounts in the supported services (del.icio.us, last.fm, flickr.com and youtube.com). This was changed when we introduced the opportunity to link to Facebook.

The event recommendation case II was the first one with the Facebook option. The four test users of the case created 7, 11, 13, and 39 interests respectively for their profiles. Two of them used the opportunity to import data from Facebook and the largest numbers of interests resulted from Facebook import. Users hesitated to connect their Facebook accounts. They were worried about the application posting messages on their Facebook wall and about their contact information spreading to advertisers. They also wanted to be in strict control of what is published on their wall.

Table 2: Interests that were given by users in two cases: the magazines and event recommendation case I

Animals	Languages	Riding
Art	**Literature**	**Sports**
Baking	Making food	Stand-up comedy
Dancing	**Music**	**Summer**
Fashion	Nature	Swimming
Movies	**Photography**	**Theatre**
Furnishing	Pilates	Traveling
Gym, exercise	**Psychology**	**Zumba**
Internet	Reading	

The test users created their profiles on the web page of our profile service. This was experienced as inconvenient because the recommendations were available on the mobile phone:

> *"I could use those interests if they were integrated into the application, so that it would not be necessary to go separately to the web page to type them in, feels laborious. And if there were enough events. We were maybe a bit too specific in our choices [of interests] and the app did not find such events."*

The recommendations shown via the HbbTV application were generated with the help of the same portable profile as for the mobile app and this helped test users to better understand the idea and value of the profile portability.

In the initial co-creation phase for the Mediatutka application, carried out in Owela, users discussed automatic data sources and updates to the profile. Users were more positive about importing their Facebook data than about letting the app analyse their browsing history. Many users expressed their concern of Facebook being able to keep track and collect information about their web browsing:

> *"I understood that I can decide myself whether I import my Facebook likes or not. I'm not worried about that. Apps that I use with Facebook login, will become 'under the control' of Facebook, and I do not like that Facebook follows me everywhere."*

> *"It is OK to log in with Facebook, but it should be possible that Facebook would not get information of what I do on other web pages."*

In the Mediatutka user test, data from Facebook was transferred automatically when the user logged in with his or her Facebook credentials. During the test, 167 Facebook interests were linked to their semantic meanings and added to the user profiles (7.3 interests/user). Half of the test users also added interests manually: 100 interests in total or 7.7 interests/user.

The interests imported from Facebook consisted mostly of proper names such as names of public persons, from politicians to artists and sportsmen, names of movies, TV programmes, actors or directors, and names of locations (Figure 3).

Person
Anton Corbijn
The Stig
Zooey Deschanel
Paolo Coelho
Barack Obama
Banksy
Lady Gaga
Chetan Bhagat

Location
Khatmandu
Pasila
Suomenlinna
Vantaa
Kuopio
Columbia Road market
Vihti

Music genre
Folk music
Blues

TV Show
The Simsons
True Blood
Seinfeld
Airwolf

Organisation
The City School
Wikileaks
TechCrunch
Wikileaks
TED
Helsinki University of Technology

Music artist
Jukka Leppilampi
Glenn Miller
Patricia Kaas

Figure 3: Examples of interests that were extracted out of the Mediatutka test users' Facebook data

3.2 Portable profile service concept

The concept of portable profiles - being able to create and manage one's own interest profile and to use it in different services when the users wants to do so - was well received by the magazine case test users (as presented in Figure 4):

- 64 % regarded the service concept as very or fairly interesting,
- 57 % regarded the service concept as very or fairly useful, and
- 47 %, considered it very or fairly probable that they would use such a service if it was available.

Free text comments revealed worries about privacy and security. Also the potential misuse of the profile, for example, for direct marketing, worried some test users. Some users also expressed concerns about how much work it would require to create and maintain the profile. Similar concerns were expressed in both event recommendation cases. These test users suggested that the application would learn their interests based on which events they have attended and which events and event categories they mostly browse. This would reduce the effort needed to create and maintain the profile.

Regarding sharing the profile, users suggested that it should be possible to offer only a selected part of the profile to a new service for making recommendations.

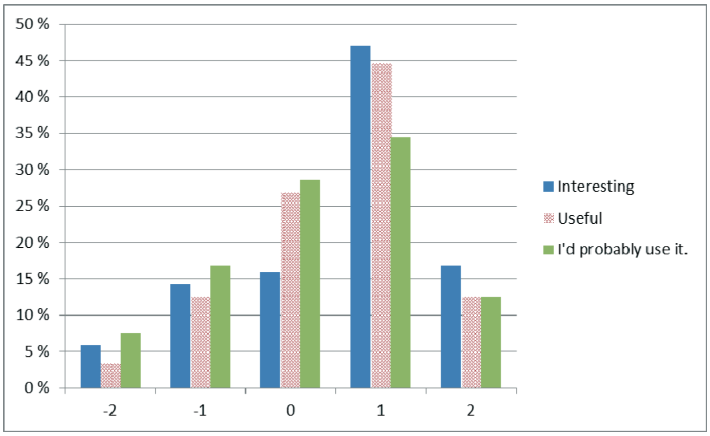


Figure 4: The opinions of 119 test users in the Magazine recommendation case regarding claims about the profile service. Main question: How do you feel about a profile that you could maintain in one service and use in different services to get personalised content recommendations. Scale: -2 = completely disagree; -1 = partly disagree; 0 = neither agree nor disagree; 1 = partly agree; 2 = completely agree

This would lower the threshold for sharing personal data with a new service. Also, in the Mediatutka co-design phase using Owela, dividing the profile into differrent levels of security was proposed as a solution to better manage the use of the service, particularly if the profile is connected to the user's real identity. Owela users were worried about their privacy and about controlling their data successfully:

"I could try a portable profile, particularly if controlling my data (profile, and use of web) has been solved well."

"In principle, using the same username and password would make life easy and ideally one would need only one set of credentials, but in practice I'm distrustful of transferring data between services. Also, of all the data that services collect about their users, often without the users realising it."

Users were worried about revealing too much information about themselves by mistake:

> *"I'm scared of the concept of a portable profile, so that more people would know more about me. Even though I would decide myself where and what will be transferred, I would at some point click the wrong button, and there goes my profile somewhere again."*

It was also suggested that the profile service should keep a detailed log of who had accessed the user data and when. The possibility to remove all personal data from the service was also mentioned as an important requirement.

3.3 Recommendation based service concepts

Our tests of recommendations in connection with different types of media content have revealed both some common and some application specific features and requirements.

In the magazine case, users could access the articles only via the recommendation list. In the free text comments, users suggested introducing search as well as browsing by category:

> *"It was nice to be able to read articles, but one could choose which one to read if they had been categorised somehow."*

Test users also expressed conflicting concerns about a recommendation based service: some worried that they would be faced with information overflow with all the recommendations, whereas others were concerned that they would miss some interesting content if the content was only available through personalized recommendations.

Despite these concerns, 63% of the magazine case test users experienced it to be very or fairly fun to explore articles based on their personal profiles (Figure 5). 61% of the responders found the concept of an article recommendation service to be very or fairly interesting (Figure 6).

Magazine articles and event information are of very different nature. Magazine articles are often read alone for solitary entertainment or to get some general information and new ideas on varied topics, whereas event information is mostly needed for planning future activities.

Finding information on interesting future events is a search task and an event recommendation service clearly competes with other ways of getting information.

Test users mentioned online search, traditional channels such as newspapers and posters, and hints from friends and family as important sources of event information. Also event ticketing agencies and event organisers' web pages were mentioned as important data sources.

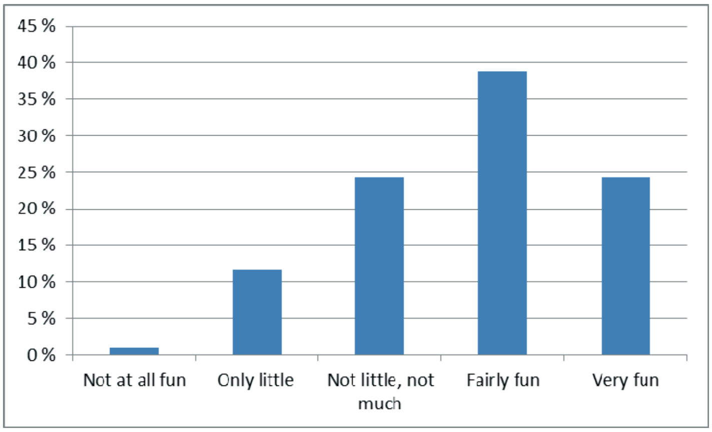


Figure 5: Answers to the question: How fun was it to explore magazine articles based on your profile? (N=103)

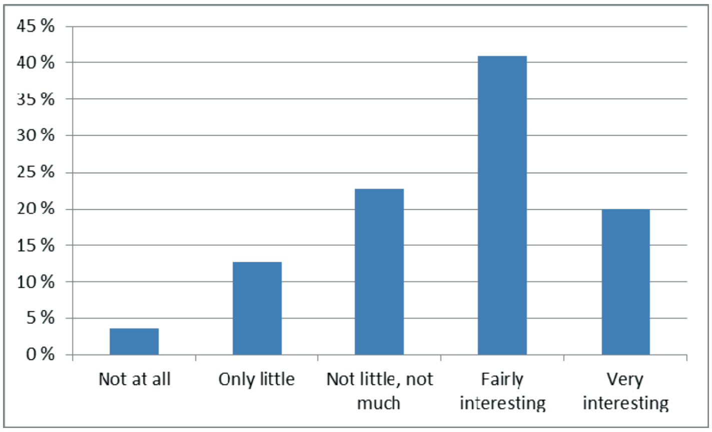


Figure 6: Answers to the question: How interesting do you find a service that recommends articles to you based on your interests? (N=110)

Even though finding event information has more the nature of a task to be performed than of entertainment, based on the brief experience with the TV application for event recommendations in case II, the test users saw event information as potentially entertaining. If and when high quality, professional videos relating to events were available, it could be relaxing to browse them at home through the TV:

> *"This is actually pretty handy, looks practical and it would actually be pretty handy to search, for example weekend activities."*

Similarly to in the magazine case, event case test users did not want to rely only on recommendations, but expressed the need to be able to search and filter events actively by themselves. In the event case II, there was already a location filter but the users wanted to have more opportunities than that for filtering.

Users proposed additional services such as alerts of future events, reminders of events that the user has expressed interest in, and information about additional services needed when going to an event.

Particularly, once the user has found an interesting event, information about various kinds of additional services and alerts are welcome. Several test users expressed their interest also in being alerted about interesting events in advance. For example, an 18-year old secondary high school student said:

> *"it would be good to have recommendations about concerts well in advance"*.

One of the test users wished that only events that match his interests really well are recommended to him. According to him, it is better to show nothing than recommendations that are only loosely related to his interests.

Since going to events is a social activity, an event recommendation application would need to be able to take into consideration the profiles of several persons. The social features of the app were highly appreciated by the test users:

> *"That "friends" view is really good."*

Deciding about going to an event with friends often requires communication and this should be easy and flexible to arrange.

Easy sharing through Facebook was suggested by one test user in case II. Our test users did not, however, wish everything be open to their friends but wanted to have the opportunity to keep some events private.

A specific requirement relating to events is that the service needs to include information of all events in the geographical area that the service specifies as its coverage. Users appreciate being able to be certain that the data covers all relevant events and that they need not look for event information from additional sources.

In the initial co-creation phase for the Mediatutka application, automatic interest based notifications were discussed. An easy way to control how recommendations and alerts are received was of high priority. As in earlier cases, users were afraid of getting a too narrow view of the world if they followed matters only via personalised recommendations. Some users even questioned the value of profile based recommendations:

> *"I see no added value. Besides, it is extremely limited if, based on the profile, the radio would every day play the same songs."*

In the actual Mediatutka user test, many users welcomed the idea of location based recommendations. Taking the recommendations where the user moves provides a new dimension to recommending content and several test users found it enjoyable. Particularly those persons who use public transport are given the opportunity to explore recommendations on the move via their mobile devices.

> *"Push notification based on my location is a good feature which I used a lot when on the move (bus, metro)."*

> *"The basic idea is good. When on the move in the city, I would be glad to keep on an app that tells me of topical things around me."*

Persons with a tight schedule and/or driving their own car or bicycle appreciate getting their recommendations and alerts just before starting a trip and not during it.

4. Discussion

User data and user profiles are key factors in web based business that offers personalised recommendations and even more so when developing personalised digital services. We have created a prototype of a profile service that lets users create and maintain their profiles for use in different media services. In our tests, users created their profiles by manually registering interests and by importing their data from various social media services.

Most user profiles in our tests consisted of 10 to 20 interests. This is a fairly small number, but still, it was enough to give users a practical experience of profile creation and utilisation in connection with recommendations. Profiles could be expanded and enriched with the help of linked data to give enough information for making recommendations. Users welcomed the idea of being able to control their data but they had many questions and doubts as to whether the data can be controlled and privacy maintained well enough also in practice. Users were afraid that they would by mistake reveal too much information. We could also see that some users wished that the profile would be updated automatically based on their actions. This indicates a willingness to let the service gather their data as long as there is clear benefit to the user.

The barrier to sharing profile data without any connection to the real identity is, of course, lower than sharing with identity, but in practice anonymity is difficult to maintain (Kobsa, 2007; Narayanan and Shmatikov, 2008.). This problem is emphasised by the increased use of mobile devices with location tracking capacities and social networking information. These types of information make anonymity an even harder goal to reach (Toch, Wang and Cranor, 2012). If the user proposed solution of sharing only a selected part of the profile will be implemented, simplicity and ease of use must be key design goals.

Users were doubtful about security particularly when data is shared between services. People's discomfort with being monitored in their web usage became clear also in connection with Facebook: several users had the impression that Facebook is tracking their web usage everywhere where the Facebook login is used and this has caused resentment. In our cases, importing interests and data from Facebook more acceptable than that Facebook would collect information on the person's web usage.

People's lack of trust in the current ways of handling user-related private data is well grounded considering the recent news of widespread data surveillance and data mining activities, not only by advertising companies but also by governments. News like this contributes to increased awareness of problems with online privacy and we can expect growing suspicion towards gathering and sharing of personal information. On the other hand, awareness of privacy issues creates a demand for systems that offer good control of personal data. A user controlled profile is only one part of the big puzzle and the entire life cycle of user data should be made transparent enough to encourage users to share their data for purposes that they regard as valuable.

The data that people provide when creating their profiles is very much determined by the way that the information is asked for. General questions produce general interests. Recommendations can be based on general interests, but in order to be able to make really matching recommendations it is necessary to know specific information about the user's interests. For example, instead of the general interest of movies, the genre of movies that the user prefers or which specific movies represent the user's preferences particularly well, are very useful. The Facebook data obtained contained much information on specific interests. Contextual information, location in our last case, offered a new dimension to take into account in the recommendations and this was well received by our test users.

The possibility of modifying the profile must be well connected to the services that use the profile. Adding and updating one's profile is most motivating in connection to a real case where the profile is being utilised. In practice, building the portable profile step by step with specific application areas in mind is a practical and feasible way to create a portable, multi-purpose profile.

A limitation of our study is that profile portability could only be partially evaluated by users since in each case the profile was only used in one service and users did not get the true experience of using the same user-controlled profile in different services. Additional research relating to portability should address this issue and should, in particular, focus on the privacy and sharing aspects.

Recommendations can be used with many types of media content and information. Depending on the type of content or information that will be recommended, requirements for the recommendations vary: the requirements on correctness and relevance of the recommendation are particularly high when items deal more with information than entertainment and when the user has a clear idea of what he or she is looking for (Liang, Lai and Ku, 2006).

The interest related data of our cases is only part of the data that is included in a larger vision of user owned data (World Economic Forum, 2013; GOV.UK, 2013) where the aim is to give the access to various kinds of transaction and consumption data as well. If and when these initiatives lead to wide-spread adoption and users are willing to share this data, there will be very rich data also for making media related recommendations. A rich user model is also a prerequisite for the next generation of user-centric recommendation systems, as envisioned by Martin et al. (2011).

The user need for having good control was strong in all our cases. It was clear in relation to the profile creation and also as to how the user can access media content. A related issue is the fear of being overwhelmed which came up in connection to magazine recommendation and notifications. The user experience of recommendation based services depends on how the whole service has been built (Knijnenburg et al., 2012). In the magazine case, the users wanted to have different ways of exploring the available content and not only to access it via their profile even though they also enjoyed this type of access. This is one indication of how the users want to be able to control their media consumption and not to be led by only the profile and recommendation algorithms.

5. Conclusion

We have here presented results of user tests of semantic profiles and semantic recommendation services in four different cases. Our cases have produced a number of findings that can be used as guidelines when developing applications that need user profiles and offer recommendations.

As expected, people are concerned about their privacy and data security and this needs to be addressed very carefully when taking the concept of a portable profile further. It is important to consider profile creation, partial profile sharing and most typical services using the profile as a whole: the profile content should support the intended services and users should feel comfortable about sharing their data. When raw data for profile creation is imported automatically, these aspects need to be taken into consideration in analysing and enriching the data. Presenting the data for the user to review and finalise is one way of giving users good control.

Recommendations are and will remain an important element in various types of media and information services. Users expect accurate recommendations, but accuracy is only one of the aspects that is needed to give a good user experience. Some users are afraid of missing something if a great deal of content is accessible only via recommendations; some are afraid of being overwhelmed by too many recommendations and notifications. Both these fears need to be taken into consideration when designing recommendation based applications.

Acknowledgements

This work was supported by TEKES, the Finnish Funding Agency for Technology and Innovation, as a part of the TIVIT Flexible Services and Next Media research programs.

References

Berners-Lee, T., 2006. *Linked Data - Design Issues.* [Online] Available at: <http://www.w3.org/DesignIssues/LinkedData.html> [Accessed 1 August 2013]

Bojars, U., Passant, A., Breslin, J. and Decker, S., 2008. Social Network and Data Portability using Semantic Web Technologies. *2nd Workshop on Social Aspects of the Web.* [Online] Available at: <http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-333/saw1.pdf> [Accessed 1 August 2013]

Friedrich, P., 2013. *Web-based co-design. Social media tools to enhance user-centred design and innovation processes.* PhD Aalto University. VTT SCIENCE 34. ISBN 978-951-38-8003-3

GOV.UK, 2013. Policy. Providing better information and protection for consumers. [Online] Available at: <https://www.gov.uk/government/policies/providing-better-information-and-protection-for-consumers/supporting-pages/personal-data> [Accessed 1 August 2013]

Heitmann, B., Kim, J. G., Passant, A., Hayes, C., and Kim, H-G., 2010. An architecture for privacy-enabled user profile portability on the web of data. *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, pp. 16-23

Kay, J. and Kummerfeld, B., 2013. Creating personalized systems that people can scrutinize and control: Drivers, principles and experience. *ACM Transactions on Interactive Intelligent Systems*, 2(4), pp. 1-42

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., and Newell, C., 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction,* 22(4-5), pp. 441-504

Kobsa, A., 2007. Privacy-Enhanced Web Personalization. In: Brusilovsky, P., Kobsa, A. and Nejdl, W., eds. *The Adaptive Web*, Lecture Notes in Computer Science. Berlin-Heidelberg: Springer Verlag. pp. 628-670

Konstantin, J. A. and Riedl, J., 2012. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction,* 22(4-5), pp. 191-123

Liang, T.-P., Lai, H.-J. and Ku, Y-C., 2006. Personalized Content Recommendation and User Satisfaction: Theoretical Synthesis and Empirical Findings. *Journal of Management Information Systems,* 23(3), pp. 45-70

Martin, F. J., Donaldson, J., Ashenfelter, A., Torrens, M., and Hangartner, R., 2011. The Big Promise of Recommender Systems. *AI Magazine*, 33(3), pp 19-27

Narayanan, A. and Shmatikov, V., 2008. Robust De-anonymization of Large Sparse Datasets (How To Break Anonymity of the Netflix Prize Dataset). *Proc. of the 29th IEEE Symposium on Security and Privacy*, Oakland:IEEE Computer Society. pp. 111-125

Orlandi, F., Breslin, J., and Passant, A., 2012. Aggregated, Interoperable and Multi-Domain User Profiles for the Social Web. *I-SEMANTICS '12 Proceedings of the 8th International Conference on Semantic Systems*, pp. 41-48

Project VRM (n.d.). [Online] Available at: <http://cyber.law.harvard.edu/projectvrm/Main_Page> [Accessed 1 August 2013]

Toch, E., Wang, Y., Cranor, L.F., 2012. Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction,* 22(4-5), pp. 203-220

World Economic Forum, 2011.*The Emergence of a New Asset Class*. [Online] Available at: <http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf> [Accessed 1 August 2013]

World Economic Forum, 2013. *Unlocking the Value of Personal Data: From Collection to Usage.* [Online] Available at <http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf> [Accessed 1 August 2013]

[1] http://xmlns.com/foaf/0.1/

[2] http: //www. w3.org/Submission/siocspec/

[3] http://profile.vtt.fi/

[4] http:/oauth.net/>

[5] http://en.wikipedia.org/wiki/Representational_state_transfer

[6] http://www.json.org/

[7] http://www.w3.org/XML/

[8] http://www.w3.org/RDF/

[9] http://www.w3.org/2006/vcard/ns#

[10] http: //www. w3.org/2003/01/geo/wgs84_pos#

[11] http://purl.org/stuff/rev#

[12] http://www.holygoat.co.uk/owl/redwood/0.1/tags/

[13] http: //scot-project.org/scot/ns#

[14] http: //dublincore.org/documents/dces /

[15] http://dublincore. org/documents/ dcmi-terms/

[16] http://lov.okfn.org/dataset/lov/details/vocabulary_moat.html

[17] http: //www.w3.org/TR/2009/REC-skos-reference-20090818/

[18] http://www.w3.org/TR/uri-clarification

[19] http://onki.fi/

[20] http:// www.free-base.com/

[21] http://www. geo-names.org/

[22] http://evenemax.fi/

[23] http://owela.fi/

[24] http://www. skimm.tv/

[25] http://stadi.tv

[26] http://www. helmet.fi

JPMTR 024 | 1315
UDC 004.42:81'37

Research paper
Received: 2013-08-19
Accepted: 2013-11-01

Knowledge-based recommendations of media content - case magazine articles

Sari Vainikainen, Magnus Melin, Caj Södergård

VTT Technical Research Centre of Finland
Vuorimiehentie 3
P. O. Box 1000
FI-02044 VTT, Finland

E-mails: sari.vainikainen@vtt.fi
magnus.melin@vtt.fi
caj.sodergard@vtt.fi

Abstract

A successful media service must ensure that its content grabs the attention of the audience. Recommendations are a central way to gain attention. The drawback of current collaborative and content-based recommendation systems is their shallow understanding of the user and the content.

In this work, we propose recommenders with a deep semantic knowledge of both user and content. We express this knowledge with the tools of semantic web and linked data, making it possible to capture multilingual knowledge and to infer additional user interests and content meanings. In addition, linked data allows knowledge to be automatically derived from various sources with minimal user input. We apply our methods on magazine articles and show, in a user test with 119 participants, that semantic methods generate relevant recommendations. Semantic methods are especially strong when there is little initial information about the user and the content. We also show how user modelling can help avoiding the recommendation of unsuitable items.

Keywords: recommendation systems, personalization, semantics, semantic web, linked data, media services, metadata, user profiles, ontology

1. Introduction

1.1 Background and objectives

Media services compete for the attention of the users. As pointed out in the theory of attention economy (Simon, 1971; Davenport and Beck, 2002), human attention is a scarce commodity that determines what enters our awareness and what we decide to act on. To be successful, a media service must ensure that its content grabs the attention of the audience. One way to do this is to offer content that the user finds relevant in his/her current situation. Well-known examples from the web search domain are Google's PageRank algorithm (Brin and Page, 1998) that determines the presentation order of the search results and the AdSense program linking advertisements to search queries.

Recommendations are a central way to gain attention in digital media services. A recommender or recommendation system (RS) is a computer-aided tool that helps users find relevant items such as magazine articles, books, songs, movies, suppliers or people. An RS should assist users in avoiding poor decisions. The information that forms the basis for the recommendation system may be collected explicitly (typically from registration or profile data, users' ratings, social networks) and implicitly (typically by monitoring user behaviour, like web pages browsed) (Bobadilla et al., 2013). Recommendation can be viewed as *filtering* of data. Filtering methods are usually divided into three main categories (Meymandpour and Davis, 2013): *collaborative filtering* that employs user ratings and browsing history without further information on the items, *content-based filtering* that uses item descriptions and compares them to user profiles, and *knowledge--based filtering* that matches knowledge about user interests and preferences with knowledge about items. Knowledge is expressed as structures, such as ontologies (Middleton, Shadbolt and De Roure, 2004), restrictions and cases. In some taxonomies, the term *demographic filtering* is used to depict that common personal attributes (sex, age, profession, etc.) influence the recommendations. *Social methods* taking into consideration a user's social networks are becoming more and more important. *Hybrid approaches* integrate features from several methods. They have been found to produce better results than any single method alone, like in the Netflix Prize competition (Bell, Koren and Volinsky, 2008). One drawback in current collaborative and content-based approaches is the shallow knowledge of user interests and meanings in content items. This weakness is especially severe when new users enter a recommendation service - this is the so called cold start problem. By acquiring knowledge about the user and storing it into a user profile, recommendations can be generated right from the start. To be able to express deep knowledge about user and content, we, like some others in the field, use semantic web technologies and linked data. This makes it possible to capture multilingual knowledge and to acquire additional knowledge from external databases with minimal user input. Many semantic databases have made their knowledge available through open API's (Application Programming Interface) and use the linked data principles defined by Berners-Lee (2006). Based on these principles, Uniform Resource Identifiers (URIs) are used as pointers to concept definitions. The same concept (e.g., temperature) may be present with various naming (e.g., Fahrenheit, Celsius) in different databases but, as long as the databases include links to the same concept, the ambiguity can be resolved. The semantic web defines knowledge resources as subject-predicate-object triplets using the Resource Description Framework (RDF)[1] and the Web Ontology Language (OWL). These resources can be queried from RDF databases using SPARQL queries[2]. Large open knowledge bases include a Google-owned community-built database called Freebase[3], the Wikipedia based DBpedia[4], the geographical database Geonames[5], and the Finnish KOKO ontology of the Finnish Ontology Library Service ONKI[6]. These databases overlap to some extent, but each of them contains unique knowledge and, to maximise the covered concepts, several databases need to be used. KOKO is good for general concepts and gives information about related concepts, whereas Freebase and DBpedia contain a large amount of knowledge about persons, music, and movies. Our aim is to develop recommendations that rely on deep understanding of both the user and the content and in this way to improve the recommendation relevancy. The recommender must allow for almost automatic knowledge capture, minimizing the amount of user input necessary. In this paper, we report the results of applying these methods in recommending magazine articles. The recommendation quality was evaluated in a user test.

1.2 Related work

Meymandpour and Davis (2012) have developed linked data based similarity metrics for recommending closely related resources. Their metrics is based on shared concept features and information theory. In our work, we also match the semantic meaning of content and user profiles, but we use semantic reasoning instead of mathematical similarity measures.

Middeton, Shadbolt and De Roure (2004) take an ontological approach to user profiling for recommending on-line academic research papers and claim a performance improvement compared to non-semantic recommenders. Unlike our work, where knowledge about the user is captured from external sources, they rely entirely on monitoring the user behaviour in the actual service.

Safoury and Salah (2013) propose a recommender based on user demographics as a way to avoid the cold start problem of content-based and *collaborative filtering*. Applying the method on MovieLens[7] data, they found that the demographic data in the MovieLens dataset did not influence differentially on users' ratings. In our work, we use more knowledge of the user than just demographics. Fernandez-Tobias et al. (2011) have developed semantic-based cross-domain recommendations. Their aim is to recommend music that is relevant to a particular place. Linked data (DBpedia) is used for finding semantic relations between places and music and a weighted graph is generated to match items between the target and source domains.

In comparison, we use several linked data sources to get additional information of different domains, e.g. music, books and movies, to recommend content items based on the user interests on various levels (e.g., genre/artist name, movie/actor). The remainder of this paper is structured as follows. First, we lay out our framework with its central principle of semantic enrichment. We then present our knowledge-based recommendation methods. In the results part, we describe a user trial with magazine article recommendations and evaluate the performance of our methods.

2. Framework and methods

2.1 Semantic portable profile platform

The user profile is central in our knowledge oriented approach, as pointed out above. Our Semantic Portable Profile Platform (SP3) supports creating, managing and utilising semantic user profiles (Figure 1). The SP3 platform has been used in a multitude of applications and its methods have continuously been developed. The platform contains tools for linking interests, context and content to semantic metadata (see Section 2.2) as well as methods for generating recommendations (Section 2.3). In this article we concentrate on using semantic enrichment for recommending magazine articles. The profile portability aspects of the SP3 platform are not within the scope of this article.

SP3 lets users create and maintain their profiles by importing data from their social media accounts and by ma-

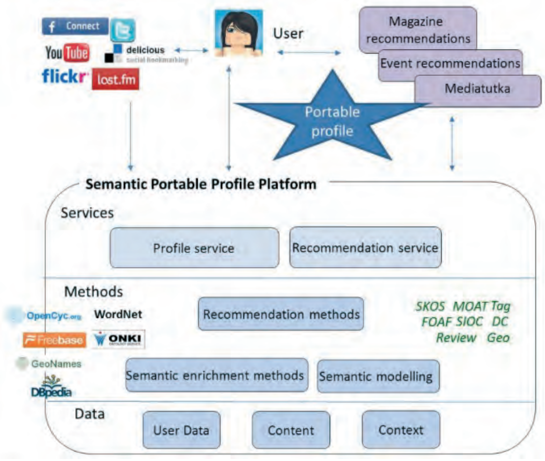


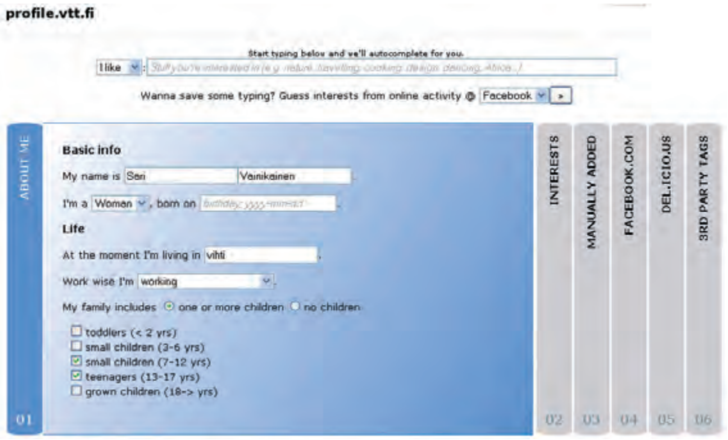
Figure 1: Semantic Portable Profile Platform (SP3). (Mediatutka refers to a mobile application which supports location based recommendations.)

nually inputting interests with the help of a semantic autocomplete widget (Figure 2). Life situation information such as family and employment situation can also be added.

The profiles are *portable,* which means that they can be used in several services. Third parties that wish to utilize the user's existing profile can do so with the user's acceptance using OAuth[8] - an open standard for authorization. REST (Representional State Transfer) APIs are offered to create, update, read and delete profiles from other services.

User and content metadata is modelled in RDF format, reusing concepts and properties from existing ontologies. The user data is coded with the help of following ontologies: FOAF[9] is used for describing demographics and online accounts, vcard[10] for home and work addresses, geo[11] for location co-ordinates, Review[12] for user ratings of interests with the values hate, like, and love, tags[13] for user interests, dcterms[14] for links to URIs of interests, and SCOT[15] for recording the account from where an interest is generated.

Content is modelled using Dublin Core dc[16] (e.g., title, description, subjects as keywords), dcterm (subjects as URIs, issue date) and prism[17] (name and number of publication, location).



profile.vtt.fi

Start typing below and we'll autocomplete for you.

I like ▾ : Stuff you're interested in (e.g. nature, travelling, cooking, design, dancing, Africa...)

Wanna save some typing? Guess interests from online activity @ Facebook ▾ »

Basic info

My name is Sari Vainikainen.

I'm a Woman ▾, born on birthday: yyyy-mm-dd.

Life

At the moment I'm living in vihti.

Work wise I'm working ▾.

My family includes ● one or more children ○ no children

- □ toddlers (< 2 yrs)
- □ small children (3-6 yrs)
- ☑ small children (7-12 yrs)
- ☑ teenagers (13-17 yrs)
- □ grown children (18-> yrs)

Figure 2: User profile creation in the profile service. Demographic and life situation information and user interests retrieved from different data sources can be viewed under the different tabs

2.2 Semantic enrichment

2.2.1 Enrichment methods

Our semantic enrichment methods use linked data for attaching metadata to users' interests or content metadata. Semantically enriched user profiles and content metadata are then used for personalized recommendations.

The workflow of our semantic enrichment methods (Figure 3) is described in the following sections.

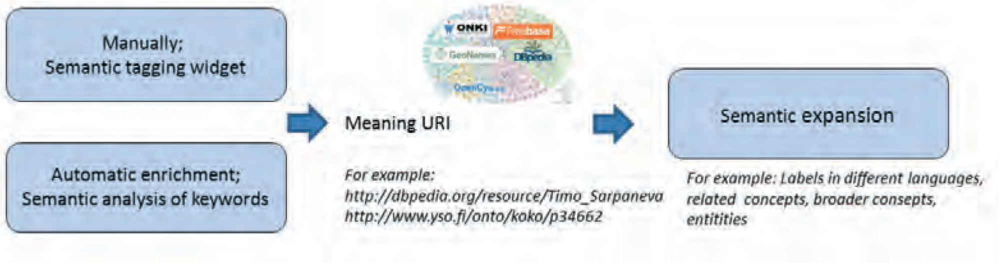

Manually;
Semantic tagging widget
Automatic enrichment;
Semantic analysis of keywords
ONKI
Freebase
GeoNames
DBpedia
OpenCyc
Meaning URI
For example:
http://dbpedia.org/resource/Timo_Sarpaneva
http://www.yso.fi/onto/koko/p34662
Semantic expansion
For example: Labels in different languages,
related concepts, broader consepts,
entitities

Figure 3: The workflow of semantic enrichment

2.2.2 Semantic tagging widget

The Semantic tagging widget (Vainikainen, Näkki and Bäck, 2012) lets the users manually add semantic tags for describing their interests. The same widget can also be used for content annotation. This autocomplete widget can be configured to use one or several linked data sources depending on the requirements of each particular case. When the user selects the suggested tag, its meaning is captured as a Linked Open Data URI.

In the profile service, automatic suggestions have been limited to general terms from the Finnish KOKO-ontology; music, movies, books, theatre and sports from Freebase; and places from Geonames. KOKO supports the management of general concepts in Finnish, Swedish and English. With Freebase, suggestions can be expanded to persons and their creations. When automatic suggestions are limited to fewer semantic databases and categories, automatic suggestions become clearer and suggesting the same concept from different semantic databases can be avoided.

2.2.3 Semantic analysis of keywords

Our semantic keyword analysis (Nummiaho, Vainikainen and Melin, 2010) uses publicly available knowledge bases - WordNet[18], KOKO, Geonames, DBpedia and Freebase - to analyse tags and keywords and turns them into semantic elements. We use it in the semantic analysis of short texts (such as TV synopses or tweets), and for analysing the keywords and tags coming from a user's social media accounts. Also service providers' vocabularies and categories are semantically enriched using this method.

The process of adding semantic meaning to keywords consists of several steps. First, we try to detect the language used. If the language is identified as English, Finnish or Swedish, we proceed to semantic annotation.

a) If the keyword is in English, we look it up in WordNet and determine its most likely meaning by finding similarities with the other keywords that it was used in conjunction with.

b) If the detected language is Finnish or Swedish, we first try to figure out its meaning using KOKO. Finnish words are POS (Part-Of-Speech)-tagged, which means that they are identified as nouns, verbs, adjectives, etc. If there is no direct match in KOKO, we try to find a match for the word's plural form, and if this is not successful, we check the spelling suggestions for the word and their plural forms. If there is no match in KOKO, we translate the word to English and look it up in WordNet.

c) For English and Swedish words without match in WordNet, we first translate the word into Finnish and look up meanings for it in KOKO, also trying the word's plural form if needed. In case we still do not find a match, we look up the word in DBPedia, Freebase, and Geonames and choose the one with highest confidence. Confidence is estimated using the Jaro-Winkler distance (Winkler, 1990).

If the language was not detected as English, Finnish or Swedish, we modify the word using spelling suggestions for English and Finnish in order to see if the word would, after all, be in Finnish or English, and do the same tests as described above.

We obtain the plural forms of Finnish words from Joukahainen[19] and Finnish spelling suggestions from Tmispell[20]. We use Suomi-Malaga[21] to POS-tag Finnish words. We get the English spelling suggestions from ASpell[22] and translations from Microsoft Translator[23]. We store all alternative meanings for the analysed concepts in RDF format using SKOS[24], MOAT[25], SCOT and Tags ontology. The primary meaning is linked to the analysed concept with the skos:closeMatch property.

2.2.4 Semantic expansion of concepts

Once the semantic meanings and their URIs have been defined, additional information relating to the concepts will be retrieved from the original linked datasets and stored in the SP3 platform databases. We combine data from the different linked data sets and load it into a common schema. The schema is based on the SKOS ontology and it defines the concepts and their relations. Additional properties will be used to define location data. The integration and simplification of heterogeneous data enables us to use it efficiently in generating the recommendations.

For every meaning (skos:Concept), we store its language versions (skos:prefLabel), its type (rdf:type) and links to similar concepts (owl:sameAs, skos:closeMatch). Relations between concepts are stored in skos:narrower, skos:broader, skos:related, skos:narrowerTransitive and skos:broaderTransitive properties. We retrieve additional information depending on the type of the concept. For example, if the concept is a music artist, information relating to music genre, bands, and other artists in the band will be retrieved; if the concept is a movie, information relating to its actors, directors, writers, genre and other movies in the same genre will be retrieved. For sports related concepts, additional information about the league, teams, players or athletes will be retrieved. For Geonames location concepts, geo coordinates and place hierarchy, such as continent, country, administrative divisions and nearby places, are retrieved. In addition to the SKOS ontology, Geonames ontology[26] properties such as featureClass, featureCode, inCountry, countryCode as well as the geo ontology properties geo:lat and geo:long are used to store location information into the profile database. We use an OpenLink Virtuoso RDF[27] database to store the semantic data.

2.3 Knowledge-based recommendation methods

2.3.1 Algorithms and workflow

Knowledge-based recommendation algorithms compare semantic content items with semantic user preferences and select the closest matches to recommendations.

We create the semantic representations using the above described methods. Mathematically, the user i gets an ordered list of content item recommendations $\{A'_{i,1},\ldots..A'_{i,p}\}$ from the matching function M having as arguments the user profiles U_i and the content items A_j for positive integers I [Equation 1]. The matching function value is called *rank value r*.

$$\forall j \in \{1,..p-1\},\; M(U_i,A'_{i,j}) \geq M(U_i,A'_{i,j+1}) \;\vee\; \forall s \in I,\, M(U_i,A'_{i,p}) \geq M(U_i,A'_{i,p+s}) \qquad [1]$$

We have developed two variants of knowledge-based recommendations representing increasing amounts of semantic enrichment and knowledge about the user: semantic and life situation specific.

As a baseline method, we use free text indexing without semantics in the content metadata creation.

The workflow of the recommendation methods is depicted in Figure 3 and described in the following sections. It is important to point out that we generate the recommendations based on the users' semantic profiles but we use additional criteria, such as publication date or location, to make the final decision on which items to show as top recommendations to the end users.

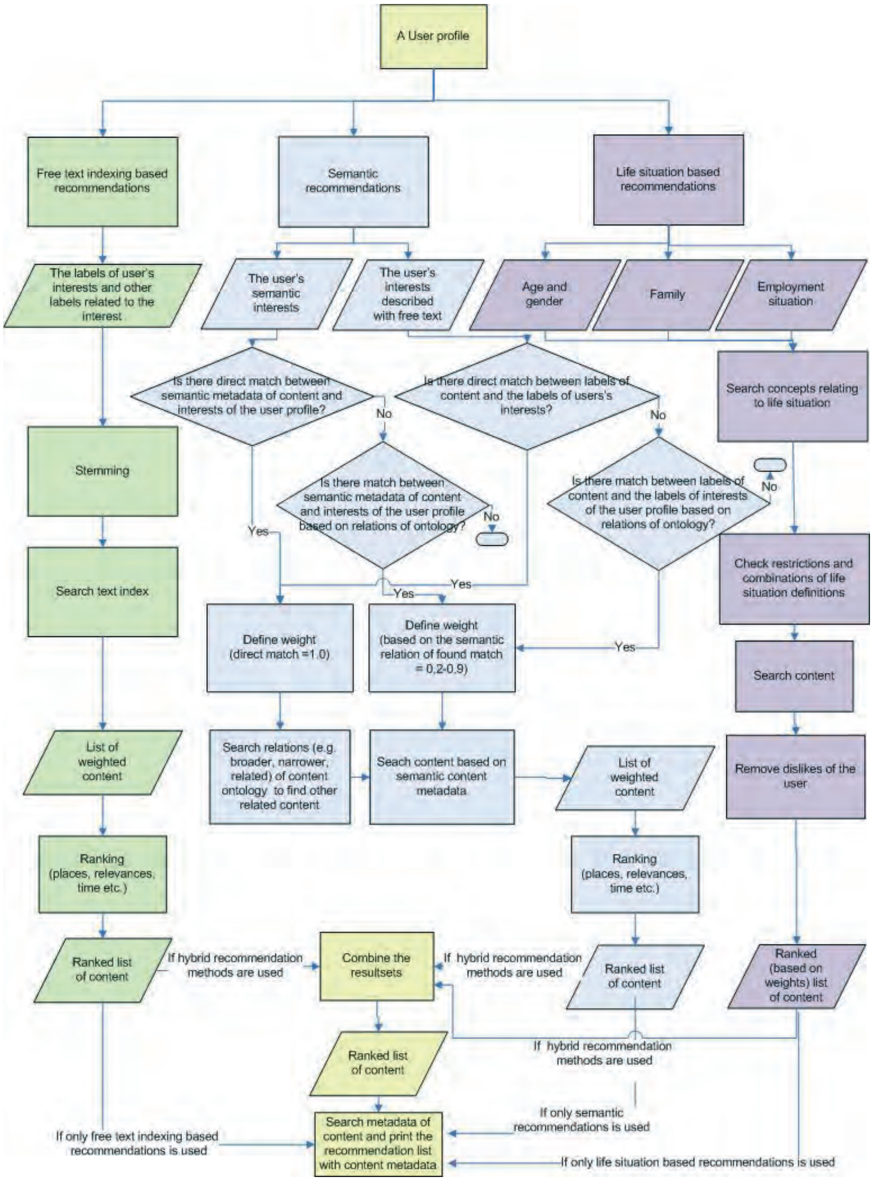

A User profile
Free text indexing based recommendations
Semantic recommendations
Life situation based recommendations
The labels of user's interests and other labels related to the interest
The user's semantic interests
The user's interests described with free text
Age and gender
Family
Employment situation
Is there direct match between semantic metadata of content and interests of the user profile?
Is there direct match between labels of content and the labels of users's interests?
No
No
Search concepts relating to life situation
Stemming
Is there match between semantic metadata of content and interests of the user profile based on relations of ontology?
Is there match between labels of content and the labels of interests of the user profile based on relations of ontology?
No
No
Yes
Check restrictions and combinations of life situation definitions
Search text index
Yes
Yes
Define weight (direct match =1.0)
Define weight (based on the semantic relation of found match = 0,2-0,9)
Yes
Search content
List of weighted content
Search relations (e.g. broader, narrower, related) of content ontology to find other related content
Seach content based on semantic content metadata
List of weighted content
Remove dislikes of the user
Ranking (places, relevances, time etc.)
Ranking (places, relevances, time etc.)
Ranked list of content
If hybrid recommendation methods are used
Combine the resultsets
If hybrid recommendation methods are used
Ranked list of content
Ranked (based on weights) list of content
Ranked list of content
If hybrid recommendation methods are used
If only semantic recommendations is used
If only free text indexing based recommendations is used
Search metadata of content and print the recommendation list with content metadata
If only life situation based recommendations is used

Figure 3: The workflow of the free text based, semantic and life situation based recommendation methods

2.3.2 Free text indexing based recommendations

Our free text indexing method matches semantically enriched user profiles with the content text index without any metadata enrichment. The semantic enrichment of user profiles gives support for multiple languages and enables extending searches to related subjects. We have used the Lucene[28] engine to index article texts and their existing metadata. Terms were stemmed in indexing.

The matching used the Solr[29] search engine which gives relevance values normalized between 0.0-1.0. These values were used for ranking recommendation results according to Equation [1].

2.3.3 Semantic recommendations

In this method, both users' interest profiles and content metadata were semantically enriched as described in section 2.2.

We used SPARQL queries to match semantic user profiles and content metadata. Ontological concepts may match directly or via closer or more distant concept relations. Weights were defined based on the distance of the match, so that a shorter distance gave a higher weight.

When an exact match was found between an article and a user interest, ontology relations were used for searching additional content to recommend. Weighted distances and the user-defined level of user interests (hate, like, love) were used for calculating the rank values. When several interests of the user profile match, the rank value increases.

2.3.4 Life situation based recommendations

In this method, additional knowledge was added to the user profile. We studied how to model the user's life situation and how to take it into consideration in recommendations. Life situation definitions were based on the user's age and gender, family situation such as the age of children and the employment situation, such as working, student or retired (Table 1).

Each life situation is described by an ontology with a certain concepts set. First, we looked up all concept URIs that matched each life situation that the user belonged to, meaning gender and age, employment situation, and family situation. We removed from this set of URIs all concepts that, while relevant, should not be recommended (e.g., "how to prevent burn-out at work" for the unemployed). The resulting set of URI's represents the user's stereotypical interests.

After this, we constructed SPARQL queries to find articles that match each of these stereotypical interests. Known user dislikes were filtered out of the results. We also searched for articles that would match skos:narrowerTransitive or skos:narrower of the concepts URIs of the defined life situation. The resulting articles were given weights depending on which hierarchic relation was used to obtain the match.

Table 1: Overview of the life situation definitions used in our work

Life situation		The ontology used for modelling	Examples

Age and gender	12 different groups based on gender and age limits	FOAF
		foaf:Group, foaf:gender,
		foaf: topic interest
		owl:Restriction for age limits
			women_15to19;
			movies, sport, music, shopping, fashion, make-up and trends.
			women_50to64;
			food and drink, travelling, handicrafts, culture and nutrition.

Employment situation	entrepreneur, working, unemployed, student, a stay at home mom/dad, and retired	SKOS owl:NegativePropertyAssertion for defining which concepts should not be recommended	working; wages, work, travelling to work… unemployed; unemployment, job interview...
Family situation	Number of children (one or more children, no children) Ages of the children	FOAF foaf:Group, foaf:topic_interest owl:NegativePropertyAssertion for defining which concepts should not be recommended together	toddlers (<=2 years) small children (3-6 years) small children (7-12 years) teenagers (13-17 years) grown-up children (>=18) familyNoChild E.g. children, child diseases, parenthood, children's culture and children's clothes should not be recommended to a person with no children

2.3.5 Hybrid methods

We combined our semantic and text index based recommendations by calculating normalized rank values of the two recommendation result lists. The items that were found in only one of the lists were given the rank value of this list, and the items that were present in both lists were assigned a rank value that combined the two rank values. We used the following formula in combining the rank values: $r = max(r_1, r_2) + sqrt(min(r_1, r_2))$, where r_1 is the semantic recommendation rank value, r_2 the text based recommendation rank value, and r the final rank value of the recommendation. We also tested linear ($r = r_1 + r_2$) and squared ($r = max(r_1, r_2) + min(r_1, r_2)$^2) methods for combining the rank values, but the square root approach produced the best results.

3. User tests

We tested the quality of recommendations with actual users and 606 magazine articles from 12 women's magazines published by Sanoma Magazines Finland as the test set. Users were recruited by email. In total, 337 users created a profile and 119 out of them completed the entire user study.

The age distribution was between 18 and 64 with 58% of the participants being under 35 years of age. Most of the participants were women (116/119). Four recommendation methods - free text based indexing, basic and modified semantic as well as life situation semantic - were developed and tested (Table 2).

Table 2: Recommendation methods and their testing

Method	Content and logic	Profile	Testing
Basic semantic	Manually created metadata using the magazine vocabulary linked to semantic databases such as KOKO, Freebase and DBpedia to obtain semantic meanings of the concepts.	Semantically enriched concepts.	Online in two phases. 1. Users created profiles (interests, dreams, etc.) manually either by using free words or by choosing words that were linked to a semantic database. Linking to social networks (YouTube, etc.) was offered. 2. Users rated the recommendations
Free text indexing based	Text index of articles	Same as above	Same data as above
Modified semantic	Magazine vocabulary and recommendation logic were updated based on the experiences of previous tests.	Same as above	Same data as above
Life situation semantic	Magazine vocabulary ontology extended. Recommendation logic developed.	The profile page for inputting life situation data was developed	23 life situations modelled in software and recommendations subjectively tested.

Magazine articles on various topics and of various story types were chosen and a Sanoma Magazines representative added metadata manually using a magazine vocabulary[30] that was at the time under development in the company. The concepts of the magazine vocabulary were automatically analysed and links to semantic databases were added to give them the semantic meaning. The results of the automatic semantic analysis were checked manually. Of the 658 defined concepts in the magazine vocabulary, 6.2% needed correcting. In addition, names of places in the metadata of travel related articles were semantically enriched using the Geonames dataset.

The articles were available online on a web site. The users were given the tasks of creating profiles for themselves and then to rate the article recommendations that were offered to them based on their profile data. In the profile creation phase, users' interests, future plans, dreams, and current problems were asked for. Users could give their input either by entering words freely or by selecting among suggested tags using the semantic tagging widget. Users could read the recommended articles on a website that was developed for the test and they were asked to rate the recommended articles on a scale from -2 (not at all relevant) to 2 (highly relevant).

337 user profiles with a total of 4892 concepts or tags (14.8 tags/user) were created; the lowest quarter entered 9 tags at the most, half of the users gave at least 13 tags, and the top quarter entered at least 19 tags, the maximum being 45 tags. We obtained 8116 article recommendation ratings from the test users.

4. Results and discussion

4.1 Evaluation and recommendations

We calculated the precision P from the complete sets[31] of *relevant* and *recommended* articles according to Equation [2]:

$$P = \frac{|\{recommended\} \cap \{relevant\}|}{|\{recommended\}|} \qquad [2]$$

In accordance with, e.g., Bobadilla et al. (2013), we consider the recommended article to be relevant, if the user gives it a rating of 1 or 2 on the 5 point scale from -2 to 2.

The precision P_r (Equation 3) is the number of relevant ratings M divided by the total number of user ratings N

$$P_r = \frac{M}{N} \qquad [3]$$

Alternatively, the precision per user is (Equation 4):

$$P_u = \frac{\sum_{i=1}^{I} P_i}{I} \qquad [4]$$

where P_i is the precision for user i, and I is the number of users.

A measure of how well the recommendations are ordered is the Pearson correlation R between the rank value r and the user rating u (Equation 5).

$$R = \frac{\sum_{i=1}^{N}(r_i-\bar{r})(u_i-\bar{u})}{\sqrt{\sum_{i=1}^{N}(r_i-\bar{r})^2}\sqrt{\sum_{i=1}^{N}(u_i-\bar{u})^2}} \qquad [5]$$

4.2 Semantic recommendations

The central idea of enriching the magazine vocabulary with linked data, especially with the concepts of the Finnish KOKO ontology, and utilising this information in recommendations, worked well. The magazine vocabulary is an high level ontology offering the main concepts that are important for the different magazines. With the help of linked data, these concepts could be extended into more detailed concepts. The mappings between the concepts of the magazine vocabulary and the concepts of linked data helped in generating recommendations for user interests that were not directly included in the magazine vocabulary. For example, a user had defined her interest in carpets with the Swedish word "mattor". Thanks to semantics, we obtained support for multiple languages and it did not matter in which language the interests had originally been defined. The magazine vocabulary did not contain the concept "carpets" but, based on the KOKO relations, the system knew that it was related to the concept of "furnishing fabrics" and was able to recommend articles relating to that subject.

However, the method also produced false recommendations. We analysed the false recommendations and made the following modifications to the basic method:

Limiting semantic extensions

An example of a false recommendation is when an article tagged with "MS disease" leads to recommending an article about "children's diseases".

The reason for this was that, in KOKO, the concept "MS disease" links to "disease" which in the recommendation process was narrowed to "children's diseases". In this case, the recommendations should not have been extended to other diseases.

When the link to the magazine vocabulary has been found with the help of extended links in the KOKO hierarchy, it means that the match to the concept of the magazine vocabulary is not an exact match. In such cases, the recommendations should not be extended to more specified concepts since this takes recommendations further away from the user's actual interest.

Refining class hierarchy, metadata and concept relations

Some concepts in the magazine vocabulary connected two different topics into one class (e.g., travelling and nature) and this caused problems in recommendations.

If a user had indicated that she was interested in "travelling", the recommendations were extended based on the hierarchy of the vocabulary and the user also received recommendations relating to nature. Several users who had more negative than positive ratings had used these concepts for expressing their interests.

Person related roles (babies, children, adolescents, women, men, siblings and relatives) had been defined on the same level of the hierarchy. It was difficult to recommend family related articles because a simple tag "family" or "children" does not reveal what in particular the user is interested in. If a user inserted the tag "family" into her profile, the recommendations included articles on topics from childbirth and babies to teenagers.

The articles should be annotated more in detail and users should be encouraged to define more precisely what aspects of "family" they are interested in.

More relations between concepts of the magazine vocabulary would be useful, for example to indicate a link between "weight control" and "diets".

Refining the mappings between KOKO and the magazine vocabulary

We found that the hierarchical relationships between terms were sometimes problematic, especially in situations where the mappings had been made between very general concepts such as "events" and "phenomena".

Mapping to the concepts of "organized events" and "cultural phenomena" would have been more appropriate and prevented some false recommendations.

The quantitative results from the user tests of the basic and modified semantic methods are presented in Table 3.

The precision of the basic method was 0.58 and, if rating 0 was seen as relevant, 0.74. The modified method was slightly more precise (0.63 and 0.77, respectively). The precision per user was 0.68 for the basic method and 0.70 for the modified method. Using the basic method, 12% of the users gave a negative average rating, i.e., their recommendation list contained more unsuccessful than successful recommendations. Using the modified method, the number of unsuccessful recommendations could be decreased. The average ratings of three users changed from negative to positive and the average ratings of 6 users changed to highly relevant.

The number of recommendations with negative ratings was smaller using the modified method. On the other hand, the added restrictions on ontology relations also reduced the number of relevant recommendations. In addition, the Pearson correlation between rank value and rating decreased from 0.23 using the basic to 0.15 using the modified method. This implies that the rank value algorithm needs more tuning.

Table 3: The relevance of the recommendations made using the semantic methods and evaluated with the following metrics: 1. Precision and correlation. 2. The percentage of the different rating values from the total amount of user given ratings for the generated recommendation list. The rating scale is from -2 (not at all relevant) to +2 (highly relevant). N is the total number of rated articles on test users recommendation lists. 3. The percentage of users having a rating average equal to one, positive but less than one, or negative

	Basic semantic (Number of rated recommendations N= 8116)	**Modified semantic** (N=4809)
Precision P_r	0.58	0.63
Precision P_u	0.68	0.70
Pearson correlation *R*	0.23	0.15
Rating	**Percentage of all ratings**	
2	29	32
1	29	30
0	16	15
-1	12	11
-2	14	12
Average rating per user	**User %** (Number of users=119)	
≥1	40	45
0 ≤ average rating < 1	47	45
<0	12	10

To understand the correlation between rank values and user ratings better than just by looking at the Pearson coefficient we assume that, if the rank value is above a threshold (=1), the item exactly matches the user's interests. If the value is below the threshold and the match has been found based on the relations of the semantic concepts, it is assumed to match less well with the user's interests (Table 4).

Table 4: The correlation between calculated rank values and user given ratings of the recommended articles. The recommendations were generated using the basic semantic recommendation method

Calculated rank value of recommended article	**User rating**	**Percentage of all ratings** (N=8116)
High rank value (r ≥1)	Positive rating (0, 1 or 2)	48
High rank value (r ≥1)	Negative rating (-1 or -2)	11
Low rank value (r < 1)	Positive rating (0, 1 or 2)	25
Low rank value (r < 1)	Negative rating (-1 or -2)	17

Table 4 shows a good overall quality of our recommendations. However, 11 % of the rated items had a high rank value but received a negative rating from the user. This may be because the subject area (e.g., family) was so wide and diverse that not all recommended articles were of interest to the user.

The numbers indicate that, while semantic relations have the potential of finding additional interesting articles to recommend, it is important to consider carefully how these deep semantic relations should be used in order to avoid going too far away from the user interests. We also analysed which tags caused a conflict between the rank values and user ratings. Family, children, nature, vacation and travelling are examples of such tags. This supports the observation mentioned above that too general terms produce false recommendations.

4.3 Comparing semantic recommendations to traditional free text indexing

To acquire further insights, we compared the results of our basic semantic method to results obtained when the content metadata was produced using free text indexing, and also to results gained by combining semantic and free text indexing (Table 5).

Table 5: Number of recommendations using different recommendation methods

	Semantic (basic)	Free text indexing	Semantic and free text indexing
Total number of recommendations for all users	33 519	35 516	49 620
Total number of recommendations for all users with user ratings	8 116	5 077	8 116
Average number of recommendations per user	252	272	375
Number of different articles that were recommended (Total 606 articles)	603	569	603
Number of different recommended articles that had at least one rating	541	456	541

The free text indexing method generated 6 % more recommendations than the semantic recommendation method. A combination of both methods produced recommendations that could not be found using single methods; thus it generated the largest number of recommendations.

There is no big difference in the average number of recommendations per user. However, some individual users had a large variation in the number of their recommendations. User profiles influence the number of recommended articles to a great extent. General profile terms such as family, home, children and food lead to a large number of text index based recommendations because these topics are very common in women's magazines.

Although the text index based method as a whole produced more recommendations than the semantic method, 32 % of users received more recommendations using the semantic method. These users had terms such as travelling, health, nature, decoration and illness in their profiles. These terms had been semantically expanded to several other terms. There were 85 articles that had not been included in any free text indexing based recommendation but had been recommended based on semantic relations. These articles were related to science, domestic appliances, food, fashion and welfare.

Relevance

We subjectively compared the relevance of recommendations produced using the two methods. We studied users with irrelevant semantic recommendations (negative average user rating) and users whose semantic recommendations were good (average user rating higher than or equal to 1.5).

Even if free text indexing increased the number of recommended articles, it also produced irrelevant recommendations and some relevant articles that had been found using the semantic method were not found at all using free text indexing. Many of the irrelevant articles at the top of the recommendation list produced by the text method were interview articles and articles relating to divorces. They ended up at the top because they often covered several topics, such as living, friends, children, family terms also and economy - popular in the users' interests. Another problem with free text indexing is the ambiguity of words: the term "renovation" (Finnish: "remontti") was used in a divorce article in the meaning of "renovation of life", whereas the user's intended meaning was "renovation of houses".

The influence of the user profile

The number of interests and the actual words used in the user profile had large influence on the results of the different recommendation methods. When there were only few interests in the user profile, the free text indexing based method performed worse that the semantic method. This was as expected, as the latter is able to infer more terms and enrich the profile. For example, 'fashion' can be extended to different accessories and 'food' to different types of courses. Very general terms in the user's interest profile caused irrelevant recommendations using both recommendation methods. It is important to guide users to describe their interests as specifically as possible.

Strengths and weaknesses of the methods

The semantic recommendation method provided the best results when both the user profile and the content metadata were semantically enriched. The drawback is that semantic annotation, even when performed automatically, sets additional requirements on the production systems. The benefit of free text indexing is that it is easy and cost effective to create.

The best recommendation results were achieved by combining the two recommendation methods. The combination creates benefits, such as:

- Finding relevant articles that cannot be found using one method only.
- In the semantic recommendation method, free tags are matched to the concepts of the magazine vocabulary but, since the number of the defined concepts is fairly low, matches cannot be always found. Text indexing produces many descriptive terms for each article.
- When a user interest is not included in the magazine ontology, semantic relations can be used to find related subjects. For example 'golf' is not part of the magazine vocabulary but if the user has 'golf' in her interests, other sports related articles can be recommended using semantics. Using text indexing, articles containing the word 'golf' can be found as well.

4.4 Life situation semantic recommendations

When analysing the false recommendations of the basic method, we found a good opportunity for improving recommendations by modelling life situations. For instance, the terms "abortion", "childlessness" and "miscarriage" should not lead to recommendation of family related articles. The problems with these terms were caused by both the magazine vocabulary and KOKO. In the magazine vocabulary, the concept "childlessness" was linked to family-related terms. The term "miscarriage" was more complicated: in KOKO, it was related to "pregnancy" and this created relations to the "pregnancy", "family" and even "breastfeeding" concepts of the magazine vocabulary. This produced inappropriate article recommendations and gave us the idea of creating a solution for managing negation.

Using the life situation semantic method, we produced and subjectively evaluated recommendations based on 23 life situations. In defining the situations, we used the demographic classification of their readers that Sanoma Magazines uses. An example: a woman, born in 1981, is connected to the life situation 'women_30to39' and she will get recommendations about subjects such as living, nature, food and drink, children, family, and sport. If she adds to her profile that she has children under 2 years, she will get recommendations relating to babies, and the recommendations will not any more include articles about older children. If she tells that she works outside of home, she will additionally get recommendations of articles relating to work. When she adds some interests, such as dance, she will also get recommendations relating to these topic.

When we subjectively evaluated the results of the life situation based recommendations, we concluded that:

- *The life situation is especially important when there is only little information about the user.*
- *Recommendations based on incompatible life situations should be excluded*

  When analysing the recommendations for retired persons, we noticed that there were irrelevant recommendations relating to the concept "sleep" ("uni", in Finnish). The concept "sleep" was connected to the stereotypical definitions of retired persons, but a problem emerged because the result set included recommendations to articles dealing with "sleep and children".

  In order to avoid this kind of recommendations, we updated the stereotypical definition of retired persons by adding definitions that described which concepts, such as children, should not be recommended together with retirement. owl:NegativePropertyAssertion definitions were used for this purpose.

- *Combining information from different aspects of life situation is a sensitive matter*

  The employment situation definition, such as "stay _at_home_mom_dad", includes mappings to children but the family definition can contain more detailed information about the age of the user's children. In this case, recommendations should prioritise articles relating to the correct age of the children.

  The same problem occurs with life situations that are defined based on age and gender, because they include mappings to the concepts of children and family. If there is information about the user's real family situation, it should be used in recommendations.

- *Life situation definitions and user interests may cause conflicts*

  An example of possible conflicts between user's interests and the life situation is that a user had defined that childlessness is a sorrow in her life. In this case, there should not be recommendations of articles relating to children, although the user may match a life situation that includes such a definition. One opportunity for future work is to make the life situation based profiles visible to users so that they can modify and control them.

5. Conclusion

We have developed recommendation methods for media content based on knowledge about the user and the content. The knowledge is expressed using the tools of the semantic web and linked data, making it possible to capture multilingual knowledge and to infer user's interests and content metadata. Linked data also allows us to automatically retrieve knowledge from various sources with minimal user effort. In addition, *a priori* knowledge of the user solves the so called "cold start" problem that *collaborative filtering* and content-based methods encounter, because with new users these methods lack enough user interaction data to make reliable recommendations.

We have applied our methods to recommending articles from women's magazines. The user tests with 119 participants verified our hypothesis that semantic methods generate relevant recommendations. Semantic methods are especially strong in cases where there is little knowledge about the user and the content. This is because semantic ontologies make it possible to infer additional user interests and content metadata. Therefore, an exact match between user interest and content metadata is not required as is the case when using non-semantic methods.

However, the semantic methods also produced false recommendations. In analysing these, we found out that there is a need to limit the usage and influence of broader concepts. If too distant concepts are used, there is a risk that the corresponding recommendations do not anymore interest the user. Another finding was that the life situation of the user is a good addition to the user model - especially when there is only little information about a user. For women's magazines, especially information about the family situation helps in identifying relevant articles to recommend. When we added the life situation model to our methods we were able to define which concepts should not occur together (e.g. childlessness - family related articles), thus avoiding inappropriate recommendations. The life situation models developed need to be tested in more extensive user tests in the future.

The tests also show that the best recommendation results were achieved by combining the semantic and the traditional text indexing methods. This helped finding relevant articles that would not have been found using one method only. This is in congruence with results from other studies that stress the merits of hybrid methods.

Creating domain knowledge is a challenge in knowledge-based recommendation systems. We can benefit from linked data in creating such knowledge and to model it so that the same concepts will be understood in the same way in different systems. This is an important aspect also in relation to the concept of portable profiles. The use of linked data sources creates new opportunities but also challenges when developing new recommendation methods. There are many different databases, APIs and data types. The amount and accuracy of data varies among concepts and databases. The future availability of open knowledge bases is an open question but many such data bases are currently available for downloading and can be installed locally.

In our future work, we intend to extend the user model to take into account additional aspects, such as values, roles and intentions of the user. The profiles need to be made adaptive to changes in user preferences. We also plan to include context information, such as the location of the user. Context information helps in recommending relevant content considering the current situation of the user. This requires more accurate user modelling including multiple sets of interests related to different contexts and roles. Similarly, the social networks of the user are important to consider in making recommendations.

Once the user profile is available in semantic form, it can be used for personalizing a wide range of services, such as search and content adoption. We envision that portable profiles under user control will become a central element in future digital services. Privacy issues and cross-service profile visibility must be considered. To advance this vision, we have created the SP3 platform that we have used in this work.

Acknowledgements

This work was supported by TEKES, the Finnish Funding Agency for Technology and Innovation, as a part of the TIVIT Flexible Services and Next Media research programs. Special thanks go to Sanoma Magazines for co-operation in this research.

References

Bell, R., Koren, Y., Volinsky, C., 2008. *The BellKor 2008 Solution to the Netflix Prize.* [Online] Available at http://www.netflixprize.com/assets/ProgressPrize2008_BellKor.pdf> [Accessed 10 October 2013]

Berners-Lee, T., 2006. *Linked Data.* [Online] Available at <http://www.w3.org/DesignIssues/LinkedData.html>. [Accessed 10 October 2013]

Bobadilla, J., Ortega, F., Hernando, A. and Gutiérrez, A., 2013. Recommender systems survey. *Knowledge-Based Systems,* 46, pp.109-132

Brin, S. and Page, L., 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems,* 30, pp. 107-117

Davenport, T. H. and Beck, J. C., 2002. *The Attention Economy: Understanding the New Currency of Business.* Boston: Harvard Business Press

Fernández-Tobías, I., Cantador, I., Kaminskas, M. and Ricci, F., 2011. A generic semantic-based framework for cross-domain recommendation. *ACM Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems,* pp. 25-32

Meymandpour, R. and Davis, J. G., 2012. Recommendations Using Linked Data. *ACM Proceedings of the 5th Ph.D. workshop on Information and knowledge,* pp. 75-82

Middeton, S. E., Shadbolt, N. R. and De Roure, D. C., 2004. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems*, 22(1), pp. 54-88

Nummiaho, A., Vainikainen S. and Melin, M., 2010. Utilizing Linked Open Data Sources for Automatic Generation of Semantic Metadata. In: *Metadata and Semantic Research - Communications in Computer and Information Science*, Vol. 108, pp. 78-83. Berlin: Springer

Simon, H., 1971. Designing Organizations for an Information-Rich World. In: M. Greenberger, 1971. *Computers, Communication, and the Public Interest.* Baltimore, MD: The Johns Hopkins Press

Safoury L. and Salah A., 2013. Exploiting User Demographic Attributes for Solving Cold-Start Problem in Recommender System. *Lecture Notes on Software Engineering*, 1(3), pp. 303-307

Vainikainen, S., Näkki P. and Bäck A., 2012. Exploring Semantic Tagging with Tilkut. In: A. Lugmayr, H. Franssila, P. Näränen, O. Sotamaa, J. Vanhala and Z. Yu, eds. 2012. *Media in the Ubiquitous Era: Ambient, Social and Gaming Media.* Hersheu, PA: IGI Global. pp. 130-148

Winkler, W. E., 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods* (American Statistical Association), pp. 354-359

[1] http://www.w3.org/TR/PR-rdf-syntax/

[2] http://www.w3.org/TR/rdf-sparql-query/

[3] http://www.freebase.com/

[4] http://dbpedia.org/About

[5] http://www.geonames.org/

[6] http://onki.fi/

[7] http://movielens.org/

[8] http://oauth.net/

[9] http://xmlns.com/foaf/0.1/

[10] http://www.w3.org/2006/vcard/ns#

[11] http://www.w3.org/2003/01/geo/wgs84_pos#

[12] http://purl.org/stuff/rev#

[13] http://www.holygoat.co.uk/owl/redwood/0.1/tags/

[14] http://purl.org/dc/terms/

[15] http://scot-project.org/scot/ns#

[16] http://purl.org/dc/elements/1.1/

[17] http://prismstandard.org/namespaces/basic/2.1/

[18] http://wordnet.princeton.edu/

[19] http://joukahainen.puimula.org/

[20] https://github.com/voikko/tmispell

[21] https://github.com/voikko/corevoikko

[22] http://aspell.net/

[23] http://msdn.microsoft.com/en-us/library/ff512423.aspx

[24] http://www.w3.org/2004/02/skos/core#

[25] http://moat-project.org/ns#

[26] http://www.geonames.org/ontology#

[27] http://virtuoso.openlinksw.com/

[28] http://lucene.apache.org/core/

[29] http://lucene.apache.org/solr/

[30] http://onki.fi/en/browser/overview/magazine

[31] More precisely, we consider the recommendation set to be the recommendation list subset that the user has rated. Thus, the size of the set varies from user to user, the average being 48 ratings.

JPMTR 025 | 1312
UDC 004.78:004.58

Research paper
Received: 2013-07-08
Accepted: 2013-11-11

Learning user profiles in mobile news recommendation

Jon Atle Gulla[1], *Jon Espen Ingvaldsen*[1], *Arne Dag Fidjestøl*[2], *John Eirik Nilsen*[1], *Kent Robin Haugen*[1], *Xiaomeng Su*[2]

[1] Department of Computer and Information Science
Norwegian University of Science and Technology
Sem Sælands vei 7, Gløshaugen
N-7499 Trondheim, Norway

E-mail: jag@idi.ntnu.no

[2] Research and Future Studies
Telenor Group, Norway
N-1331 Fornebu, Norway

Abstract

Mobile news recommender systems help users retrieve relevant news stories from numerous news sources with minimal user interaction. The overall objective is to find ways of representing news stories, users and their relationships that allow the system to predict which news would be interesting to read for which users. Even though research shows that the quality of these recommendations depends on good user profiles, most systems have no or very simple profiles, because users are reluctant to giving explicit feedback on articles' desirability. In this paper we present a user profiling approach adopted in the SmartMedia news recommendation project. We are building a mobile news recommender app that sources news from all major Norwegian newspapers and uses a hybrid recommendation strategy to rank the news according to the users' context and interests. The user profiles in SmartMedia are built in real-time on the basis of implicit feedback from the users and contain information about the users' general interests in news categories and particular interests in events or entities. Experiments with content-based filtering show that the profiles lead to more targeted recommendations and provide an efficient way of monitoring and representing users' interests over time.

Keywords: recommender systems, personalization, Big Data, user click analysis, news apps, content-based filtering

1. Introduction and background

1.1 Rationale for this study

We have in the last few years witnessed the introduction of a number of commercial mobile news apps. These are applications that help users find relevant news from a number of news sources without going to each individual source or browsing through all the news available at each source. Due to the limitations of mobile devices in terms of screen sizes and input methods, these news apps are mostly gesture-based with news headlines or news stories compressed to fit small, buttonless displays. Some of these apps, such as Summly, Wavii and Circa, present summaries of news articles to make better use of the small screens, whereas others have introduced categories and sharing with friends to help users focus on the news interesting to them and filter out the irrelevant parts (Circa, 2013; Haugen, 2013; Yeung, 2013).

It seems tempting to introduce technology that can help the news apps recommend the news that are most likely to be of interest to their users. This requires some knowledge about the news content and the users' opinions on both particular news articles and news categories in general. Explicit signals about user desirability or interests are, however, weak and the recommender system would normally need to use implicit signals such as user click patterns to infer preferences and priorities. Moreover, since a mobile user is also situated in a particular context, located at a particular place at a particular time, we may also need to assume that her context is relevant to which articles she would deem interesting. Even though a user has a general preference for sports news, for example, she might want to know that there is a traffic accident just a few blocks ahead of her.

Recommender systems have been used extensively for music and movie recommendations as well as for product reviews in general (Schafer, Konstan and Riedl, 1999) and there are already major online stores such as Amazon that offer product recommendations as part of their services. News recommendation differs in several ways from these well-known types of recommender systems: (i) articles have short life-cycles, and freshness and location may often be as important to the user as the article's content relevance, (ii) news articles are unstructured and more complex to analyze than objects with structured properties such as product reviews or networks of friends, (iii) the volatility and unlimited reach of news lead to rapid changes in both terminologies and topics over time, (iv) serendipities, or the need for variety and unexpected news, have to be addressed, and (v) cold-start problems linked to users that have no history and news that have not yet been discovered by enough users are notorious.

This paper discusses an approach for personalizing mobile news services by means of implicitly inferred user profiles. After a review of the current state of the art of recommender systems and user click analysis in section 1, we present our SmartMedia news recommendation project in section 2 and go through the steps from logging user behavior to constructing user profiles for news recommendation. Section 3 demonstrates the use of user profiling in the news recommender app and shows how news articles are ranked differently as a user's profile is updated over time. There are many complex dependencies in news recommender systems, and some of them are discussed in section 4. Conclusions and plans for further work are laid out in section 5.

1.2 Recommender systems

Recommender systems as a scientific discipline emerged in the early nineties as a particular branch of *information filtering* (Belkin and Croft, 1992). The general idea is to define techniques for predicting user responses to a given set of options on the basis of information about the options, the users and their interdependencies. The discipline draws on research from cognitive science, information retrieval, prediction theories and management science, as well as lately from semantic web and data mining (Borge and Lorena, 2010).

Formally, the problem in news recommendation is that of estimating and ranking the evaluations of articles unknown to the user. To compute these estimates, evaluations of other articles by the same users or evaluations by other users with similar interests may be used. Following Borges and Lorena (2010), we can define a set of users U and a set of news articles A; let s be a utility function (Equation 1) that defines the evaluation of an article a for a user u:

$$s: U \times A \rightarrow V \quad [1]$$

in which V is a completely ordered set formed by non-negative values within an internal, e.g., 0 to 1 or 0 to 100. The system is to recommend an article a' that maximizes the utility function (Equation 2) for the user:

$$a' = \arg\max_{a \in A} s(u, a) \quad [2]$$

An element in U can be defined by a number of characteristics that constitute the user profile of a particular user. Similarly, elements from A may be given different characteristics, depending on what information is available about the article. For example, a news article may have characteristics such as title, category, publication date, publisher, entities and locations.

It is important to note that the utility function s is not defined in the whole space $U \times A$. Estimating or extrapolating evaluations for the blanks in this space is the goal of the recommender system itself. In doing so, a whole battery of techniques may be applied, including decision trees, Bayesian classifiers, support vector machines, singular value decomposition, neural networks, clustering and information retrieval similarity scores (Adomavicius and Tuzhilin, 2005; Rajaraman and Ullman, 2011).

Recommender systems use a number of different technologies that can be classified into two broad groups: content-based filtering and collaborative filtering.

In *content-based filtering,* a new article is recommended to a user if it exhibits important similarities with her user profile. Since the user profile is constructed on the basis of her previously read articles, the recommended articles are those that are similar to articles in which she has shown interest in the past. In a vector-based system, both the user profiles and the news articles are represented as vectors in which term frequencies indicate the prominence of topics and entities, and a simple cosine similarity score may be computed to assess an article's relevance to a user. More sophisticated methods also take advantage of semantic reasoning and domain ontologies (e.g., Cantador, Bellogin and Castells, 2008). According to Borges and Lorena (2010), content-based filtering methods are effective at recommending unrated news articles, though the methods find it difficult to analyze the quality of articles or to recommend new or surprising stories that are not encoded in the user's reading history (serendipitous recommendations).

C*ollaborative filtering* can be seen as an automation of *word-of-mouth* recommendation. The idea is to recommend news articles to a user if they have been well evaluated in the past by people with similar preferences as the user. The approach can be further categorized into two types, memory-based and model-based, both of which are dependent on efficient techniques for grouping similar users together.

Compared to content-based filtering, collaborative filtering is able to make serendipitous recommendations, since similar users may still read articles that are not in the current user's own history. However, collaborative filtering needs a substantial amount of data in order to be effective. There are sparsity and cold-start problems that prevent the system from recommending new and relevant articles that have no historical ratings among the network of similar users (Borges and Lorena, 2010).

Recent *hybrid filtering* approaches try to combine the best features of content-based filtering and collaborative filtering. As demonstrated on a fraction of live traffic on Google News website by Liu, Dolan and Pedersen (2010), these combined approaches may both improve the quality of the recommendations and attract more frequent visits to the news site. More details about the various types of recommender systems, including knowledge-based filtering, are found in Jannach et al. (2010). Additional techniques that make use of contacts on social networks are presented in, for example, De Francisco Morales (2012), O'Banion, Birnbaum and Hammond (2012) and Shuai, Liu and Bollen (2012).

1.3 User click analysis

Successful news recommendation requires good and updated models of users' preferences. Unfortunately, users are often reluctant to give explicit feedback on news articles that can be inspected to construct and maintain user profiles (Thurman, 2011). This leaves us with the option of analyzing user click behavior to build user profiles that are consistent with their reading history and presumably useful in recommending interesting articles in the future. Most research on user click streams comes from the web search domain. Lee, Liu, and Cho (2005) build user models on the basis of click streams to enhance personalized web search. They infer search goals from analyzing how other users in the past have used the results of the same queries and their results suggest that the goals of up to 90 % of the search queries can be identified in this manner. Speretta and Gauch (2005) analyze click logs that consist of queries and documents clicked for every query. Like Kim and Chan (2003), they use the logs to learn user profiles that contain taxonomies of concepts, in which weights indicate the strengths of the relationships. In Nasraoui et al. (2008), clustering techniques are used to summarize user sessions into clusters that may serve as user profiles for the users in questions.

Billsus and Pazzani (2000) have developed a system for interpreting implicit user feedback on news articles presented in the Daily Learner. If a user clicks on the headline of an article, they assume that there is some basic interest in the article and set an initial score of 0.8. This score is gradually increased as the user requests more pages of the story, until a final score of 1.0 is reached if all pages have been viewed.

Similarly, a skipped article is assumed to be uninteresting and is given a negative score that is subtracted from the system's prediction score for the article. All these rated articles are afterwards combined to produce a user profile that lists weighted informative words associated with each user.

Liu, Dolan and Pedersen (2010) build user profile vectors that express users' evolving interests in specific news categories. For each user, they record the distribution of clicks and associate click rates with categories on a monthly basis. This allows them to analyze the proportion of time the user spends reading about each category as well as to reflect on the development of her interests from one month to another.

Interestingly, the design of the user interface heavily influences the user profiling techniques available to the system.

The Daily Learner can use a more fine-grained analysis than Google News because their users need to go through a series of clicks to confirm their interest and read the full news story. This may encourage the introduction of more complex user interfaces, though usability studies show that users are not very happy with news apps that require too much interaction.

2. Methods

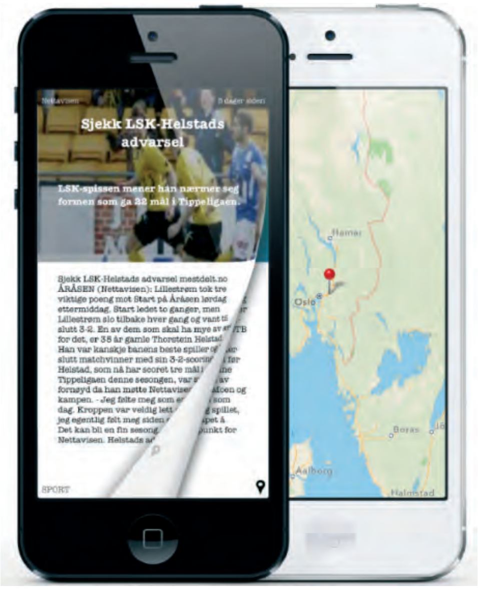
2.1 SmartMedia news recommendation

The SmartMedia project at the Norwegian University of Science and Technology (NTNU) was initiated in 2011 in close collaboration with the regional media industry and the Norwegian telecom operator Telenor Group. Central in the project is the development of an iOS[1] news recommender system app for publishing and recommending news from a number of Norwegian newspapers. An architecture based on Big Data processing pipelines and search technologies is employed to deal with the constant flow of news that is added to the SmartMedia news index. The project focus is on recommendation technologies and semantic search, making use of NTNU's experience with large-scale advanced search platforms (see, for example, Gulla, Auran and Risvik (2002), Brasethvik and Gulla (2002) or Solskinnsbakk and Gulla (2010) for the technological background of SmartMedia).

A hybrid approach to news recommendation is adopted in SmartMedia. Freshness and locational information extracted from news events and users' mobiles are part of the recommendation strategies to make sure that new events in a user's neighborhood are given sufficient attention. Addressing users' particular interests and behavior, the system combines content-based and collaborative filtering to promote news that are consistent with her previously accessed articles or preferred by other users with similar interests.

[1] Apple's operating system for mobile devices

Due to the limitations of mobile devices, the system does not assume any direct user feedback on the articles presented to her. The user will not explicitly remove or promote any articles in news streams recommended to her, as opposed to what is common in most news reader apps today (Haugen, 2013). Also, the system does not access user data from other sources that may be used to construct user profiles in the system. The only information available to the system is the observed behavior of users retrieving and reading news in the SmartMedia iOS app. Since the explicit signals about the user's interests is so weak, it has been paramount to extend the analysis of user behavior to include a broad array of complementary implicit indicators of users' interests. We consider pre-read actions such as clicking on news articles, reading characteristics such as time spent in the article view, and post-read actions such as favoriting, sharing and e-mailing article links as indicators of the user's interests. This calls for a rather complex analysis, as there are dependencies between the user actions and not all actions should contribute in the same way and to the same extent in the resulting user profile.



Nettavisen 5 dager siden

Sjekk LSK-Helstads advarsel

LSK-spissen mener han nærmer seg formen som ga 22 mål i Tippeligaen.

Sjekk LSK-Helstads advarsel mestdelt.no
ÅRÅSEN (Nettavisen): Lillestrøm tok tre
viktige poeng mot Start på Åråsen lørdag
ettermiddag. Start ledet to ganger, men
Lillestrøm slo tilbake hver gang og vant til
slutt 3-2. En av dem som skal ha mye av TB
for det, er 35 år gamle Thorstein Helstad.
Han var kanskje banens beste spiller og ble
slutt matchvinner med sin 3-2-scoring for
Helstad, som nå har scoret tre mål i ne
Tippeligaen denne sesongen, var s av
fornøyd da han møtte Nettavisen afcen og
kampen. - Jeg følte meg som e som
dag. Kroppen var veldig lett g spillet,
jeg egentlig følt meg siden pet å
Det kan bli en fin sesong punkt for
Nettavisen. Helstads ad

SPORT

Figure 1: Swiping movements are used to turn pages and navigate in news app

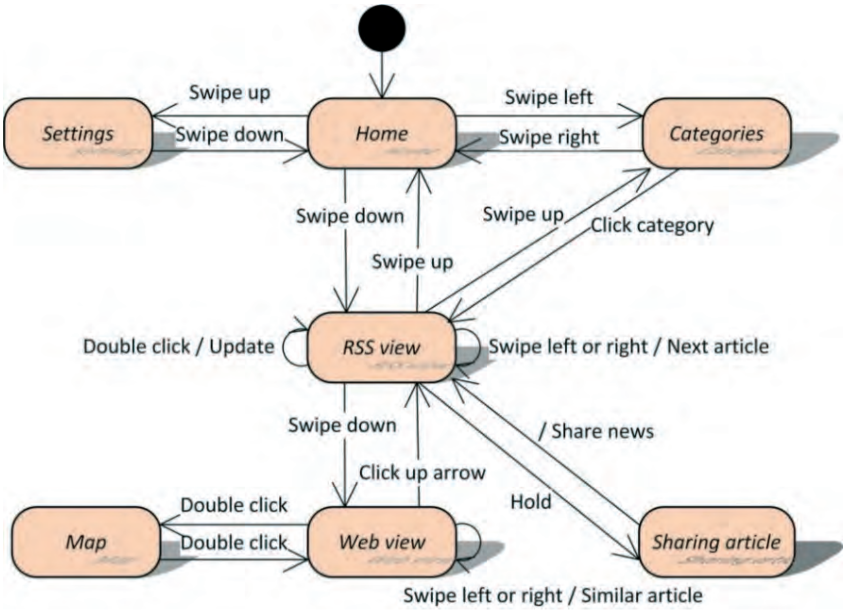


Figure 2: State transition diagram for app user interface

The SmartMedia news app has a pure gesture-based user interface as illustrated by the RSS (Rich Site Summary) and map screen shots in Figure 1. To accommodate the small screen and the lack of a proper keyboard, there are no buttons, and all navigation is done with swiping or clicking movements. The main page of the news app gives a small summary of the latest news and allows the user to log in if she would like to share her user profile across devices. Sliding side-menus are available from the main page to configure the app or select particular news categories. The user can swipe down to the RSS view, which lists - one story per page - the news recommended to this particular user. Horizontal swipes in the RSS view moves from one recommended article to the next, whereas vertical swipes take the user up to the main page or down to the full web view of the article. In the web view, the user may swipe horizontally to access related stories or double click to get a map showing where the news took place.

The state transition diagram in Figure 2 shows how the user is using gestures to navigate from page to page in the app. For more details about the implementation of the app, the reader is referred to Mozhgan et al. (2013).

2.2 Logging user behavior

The client-server architecture of the SmartMedia news app is a combination of a standard Solr[2] index for new articles, a Hadoop[3] cluster for generating user profiles, and a MongoDB[4] for event logs and generated user profiles. This Big Data approach ensures that the system can deal with the number of user actions that need to be recorded at the client side and analyzed and stored on the server side of the system.

On the server side, real-time RSS news streams from Norwegian newspapers are continuously analyzed and indexed in Solr for later recommendations. As shown in Figure 3, the indexing process accesses the RSS news before it retrieves the corresponding HTML documents. After extracting the body texts from these HTML documents, the system extracts named entities from the texts, identifies the locations of the news using Google Maps, and stores the information about every news article in a structured Solr index.

If the article is not already categorized by the publisher, a simple classifier is used to annotate the news with the appropriate news categories. Associated with every article in the index are meta-data such as publication time, geo locations, key phrases/entities, categories, and publisher.

On the client side, a middleware layer is used to retrieve ranked news articles from the Solr index and present these to the user. The user may also inspect her own user profile in case she would like to overrule or reset the profile. The iOS client logs every gesture and user click and sends these logs back to the server for storing in the MongoDB database.

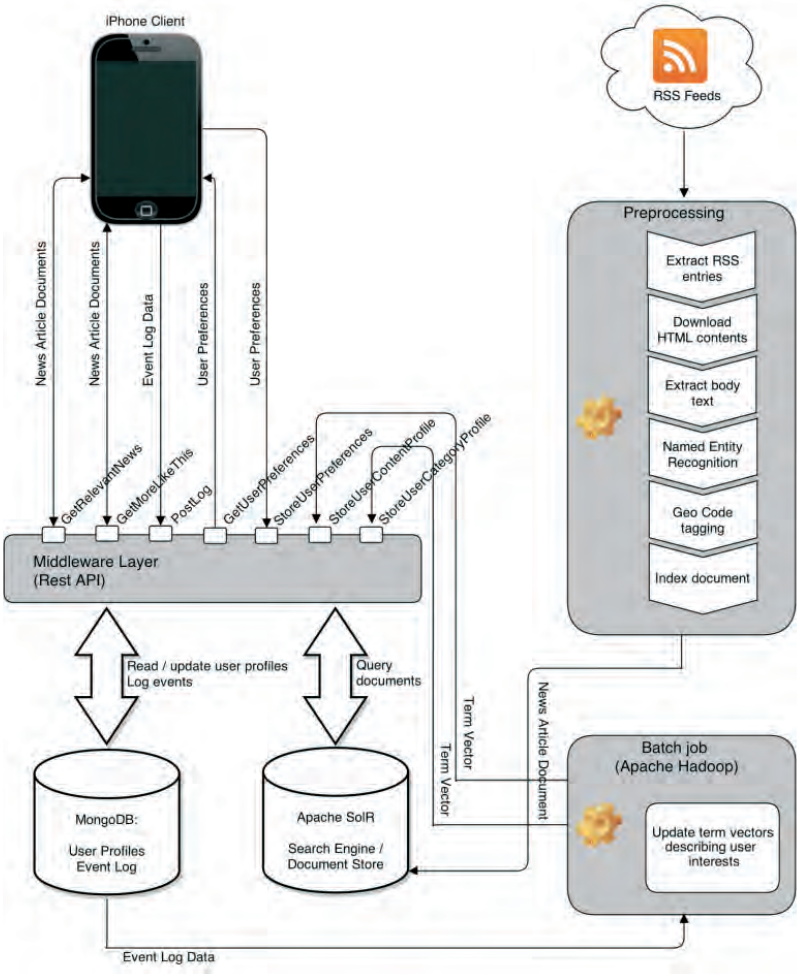


Figure 3: Architecture for user click behavior analysis

As seen in Table 1, the log information includes not only information about the user and the type of user action, but also properties of the news article itself that may be

[2] Solr is an open source enterprise search platform from the Apache Lucene projects.

[3] Hadoop is an open source software framework for processing of large data sets using clusters of hardware.

[4] MongDB is an open source document-oriented database system.

needed for the subsequent construction of user profiles. The fields 'Tags' and 'Categories' list the key-phrases and categories annotated with the article in the document index.

Table 1: Each user action is recorded as a separate record in a log database

User log data field	Description
ID	Unique identifier for the user action
User ID	ID of the user performing the action
Article ID	ID of the article subjected to the action
User Action Type	Type of user action
Timestamp	Time the action occurred
Geographical Location	Coordinates of the mobile user at the time of the action
Tags	Entities and keywords extracted from the article
Categories	Classification of article in news categories
Properties	Extra field for additional data

Since the user is not giving any explicit rating of news articles' attractiveness, the system needs to reason about her behavior to assess their value to her. The only information available, though, are the gestures and user clicks used to navigate around the user interface. On the basis of the user interface model in Figure 2, we have identified 10 user actions that may be taken as indicators of users' interests in the news article (see Table 2). For example, if the user is swiping down from the RSS view of an article to see the web view text, we assume that she has found the article interesting enough to read the full text. Similarly, we assume that she liked the article if she decides to share the article on Twitter or Facebook, store the article among her favorites, or send the article link by e-mail. If she is checking the map or requesting similar articles, it must also strengthen the view that she appreciated the article's content.

More complicated are the user actions linked to the time spent reading an RSS article or a full-text article. It seems reasonable to assume that the article is of interest to the user if she spends more time reading it than reading most other articles.

Of course, it varies from one user to another how much time is needed to read a news article. With regard to the implementation, this means that we need to monitor and store each individual user's reading habits and only consider these actions when the reading time exceeds the average reading time for this particular user.

Table 2: User actions that are used to construct user profiles

User action type	Description
Opened article view	User opened full text version of article
Time spent article view	Time the user spent viewing the article
Time spent preview	Time the user spent viewing the RSS version of article
Clicked category	User selected a news category
Shared twitter	User shared the article on Twitter
Shared facebook	User shared the article on Facebook
Shared mail	User shared the article on mail
Starred article	User added the article to favorites
Viewed map	User viewed the location of the article on a map
Viewed similar article	User accessed another article similar to current article

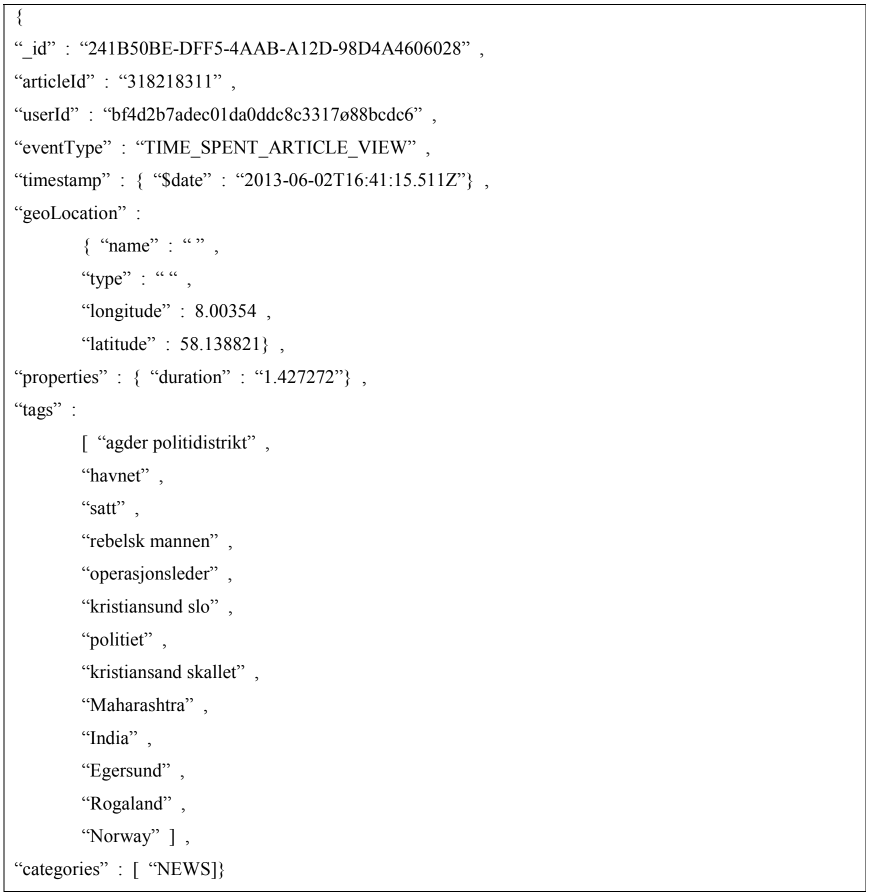
2.3 User profiles

A user profile is constructed for a chosen time interval. If there already exists a user profile for the current mobile device, the new profile is combined with the older one using a technique that renders the old profile less important than the new one.

Take the user action shown in Figure 4, which states that a user spent about 1.4 seconds reading the full text of a news article on June 2. The action, the article and the user are all given internal identifiers by the system. The event described in the article has been associated with a pair of geographical coordinates, is of the NEWS category, and seems to be dealing with an Indian man in the southwestern part of Norway that got into trouble and was imprisoned by the police.

If 1.4 seconds is longer than this user's normal reading time, the action should be taken into account when building the user profile.

The key phrases in Figure 4 point to a serious challenge in analyzing news stories. Since the terminology changes rapidly in the news domain and sentences may be both ambiguous and sometimes directly ungrammatical for literary effects, it is often difficult to extract proper key phrases and entities from stories. In this case, for example, both *'havnet'* (ended up) and *'satt'* (sat) are unsuitable as key phrases, implying that our Named Entity Recognition (NER) component suffers from a substantial amount of noise. Experiments on some articles show that almost 50 % of the phrases suggested by the NER component may in fact be noise.



```
{
“_id” : “241B50BE-DFF5-4AAB-A12D-98D4A4606028” ,
“articleId” : “318218311” ,
“userId” : “bf4d2b7adec01da0ddc8c3317ø88bcdc6” ,
“eventType” : “TIME_SPENT_ARTICLE_VIEW” ,
“timestamp” : { “$date” : “2013-06-02T16:41:15.511Z”} ,
“geoLocation” :
        { “name” : “ ” ,
        “type” : “ “ ,
        “longitude” : 8.00354 ,
        “latitude” : 58.138821} ,
“properties” : { “duration” : “1.427272”} ,
“tags” :
        [ “agder politidistrikt” ,
        “havnet” ,
        “satt” ,
        “rebelsk mannen” ,
        “operasjonsleder” ,
        “kristiansund slo” ,
        “politiet” ,
        “kristiansand skallet” ,
        “Maharashtra” ,
        “India” ,
        “Egersund” ,
        “Rogaland” ,
        “Norway” ] ,
“categories” : [ “NEWS]}
```

Figure 4: A user action recorded by the iOS client

We assume that a user's general interests can be analyzed at two levels. At the top level she might have certain preferences for particular news categories, such as sports or lifestyle news. These categories correspond roughly to the standard categories used by newspapers to structure their own content. At a lower level, a user may prefer stories about particular events, persons, companies, products, etc. These are typically found as key phrases or entities in news articles, and a particular article may refer to many of these topics with various degrees of prominence. An article may for example be mostly about the Barcelona football club but there may also be parts of it referring to Real Madrid or other Spanish or non-Spanish football clubs, as well as to persons playing for these clubs.

Our user profile is comprised of two vectors:

$$P = < \vec{C}, \vec{K} >$$

- a **category vector** $\vec{C}$ in which each news category is given a weight that indicates the importance of this category to the user,
- a **content vector** $\vec{K}$ that lists all prominent key phrases and entities that the user may find interesting to read about. The weights indicate the user's relative interest in each key phrase and entity.

Both vectors are normalized so that the maximum weight of any category or key phrase is 100. The category vector below belongs to a user that has been reading mostly news and entertainment stories.

$$\vec{C} = < (\text{"NEWS"}, 100.0), (\text{"SPORTS"}, 13.1), (\text{"TRAVELING"}, 7.5), (\text{"LIFESTYLE"}, 80.3), (\text{"ECONOMY"}, 3.14) >$$

The interpretation of $\vec{C}$ is that this user prefers straight news stories and to some extent lifestyle news, and she is not very likely to read about economic affairs. The content vector below is more difficult to interpret, as it seems that the user's interests cover a wide spectrum of news categories. Again, the terms 'i', '2', '3' and 'ST 13' should probably not have been part of the content vector.

$\vec{K}$ = < ("46354", 1.866), ("Akatsi", 1.866), ("Forbruker", 1.866), ("mortenthomassen", 0.933),
("jeg", 0.9444), ("Lodzkie", 19.633), ("Strømstad", 1.8666, ("PaysdelaLoire", 1.866), ("150", 2.833)
("RueBenjaminFranklin", 1.866), ("Nordrhein-Westfalen", 1.866), ("rachelnordtømme", 1.866), ("Nordland", 7.555),
("¡", 0.933), ("3", 2.811), ("ST 13", 19.633), ("2", 81.844),
("BestWestern Anker Hotel", 1.888),... >

The content vectors grow as users read more stories and will ultimately contain thousands of entities that the user may have found interesting at some point. However, only the higher weighted terms will be important at the recommendation stage, and the less important terms can be readily ignored or even removed from the vectors.

2.4 Automatic construction of user profiles

User profiles are constructed in two steps: (1) build a time-constrained profile that covers the time from when the last profile was generate up until the current time, and (2) merge the old profile with the time-constrained profile.

The following steps show how user *u*'s time-constrained user profile is built for the time period from t_0 to t_1:

1. *Define a user action set S that contains all user actions from t_0 to t_1 for user u*
2. *Assume a weight of 1 for all user actions in S (all actions are equally important)*
3. *Remove user actions from S about timed reading events if the time spent is less than the average time for u.*
4. *Form a category vector, in which the weight of each category represents the total number of occurrences of the category in the actions in S.*
5. *Form a content vector, in which the weight of each key phrase/entity represents the total number of occurrences of this phrase/entity in the actions in S.*
6. *Normalize category and content vectors.*

Merging the old user profile with the new time-constrained profile can be done in different ways. Intuitively, we would want to modify the old profile so that your later interests become more important than your old ones. This will make sure that irregular and important events that are unfolding right now receive enough attention, even if these topics are not necessarily found in the old user profile.

On the other hand, we need to be careful about deleting parts of the old user profile. If there are elements in the old profile that are not present in the time-constrained profile, the reason is not necessarily that the user's interests or preferences have changed. It might simply be that there has not been any news lately about those particular topics, and the user would be happy to have the topics recommended when there is news about them again.

We assume that the new time-constrained vector should be given more weight than the old user profile. There are different ways of implementing this "forget me" function, but we have adopted the following formula in our project:

$$\overrightarrow{V_u} = \overrightarrow{V_n} + c\overrightarrow{V_o} \qquad [3]$$

The new user profile $\overrightarrow{V_u}$ is given by the sum of the new time-constrained profile $\overrightarrow{V_n}$ and the multiplication of the old profile $\overrightarrow{V_o}$ with a constant with a value between 0 and 1. If *c* is set to 1, old and new user behavior count as equally important in the updated user profile. Only the latest user behavior is considered in the new user profile if *c* is set to 0.

3. Results

3.1 User profiles from interaction with news app

The SmartMedia news recommender app is already in operation and contains close to 150 000 news articles. Each day, around 1 500 articles are added from a total of 89 newspapers in Norway. The average newspaper article is 220 words long, if we exclude finance news that are usually substantially longer than news from other categories. Statistically, each article contains 1.6 location names, 2.3 person names, 2.3 organization names and 0.8 role names.

On the left-hand side of Figure 5 we show the user profile of a particular user at time t_0. The profile is automatically built from logging and analyzing all actions by the user up until time t_0. To simplify the presentation, only the top 20 terms of the content vector are shown, ordered according to their weights.

The category vector reveals that the user is using the app to read standard news stories. On the content level, we notice that there are numerous local geographical terms as well as some general terms from the latest news. Before any user profile is established, news articles are presented to the user on the basis of freshness and geographical proximity.

As the user profile is constructed, new articles are recommended and ranked according to how they match this user profile. In the current implementation, a content-based match is computed as the cosine similarity between the user profile and the vector representations of the news articles. The left-hand side of Table 3 lists the top 10 news recommended to the user at t_0. Whereas seven of the stories fall into the news category, there are also two articles about sports and one about lifestyle.

Most of the news stories concern events that take place in the vicinity of the mobile user. One explanation for this may be that the user has a strong preference for local news and has already in the past preferred such news articles. However, since geographical proximity is also used to recommend articles, her exposure to local news articles may have been so high from the outset that it was difficult not to view mostly local news.

3.2 Learning user profiles over time

Figure 5 also demonstrates the learning effect of the user profiling approach. The user profile is updated at regular intervals by combining the old profile with an analysis of what the user has read after the old profile was constructed. Whereas the user profile at time t_0 reflects her behavior up to time t_0, the new profile in t_1 for the same user is constructed from the profile in t_0 and a time-constrained profile that covers the time from t_0 to t_1. We can see that the user has gradually moved more into sports, either because she is genuinely more interested in sports or because there are relatively few proper news stories at this point. Her interests in finance news have fallen, though she seems more inclined to like travelling news now than in the past.

Time t_0			*Time t_1*	
Category C_0			Category C_1	
NEWS	*100.0*		*NEWS*	*100.0*
SPORTS	*1.47*		*SPORTS*	*27.2*
LIFESTYLE	*4.41*	→	*LIFESTYLE*	*2.91*
TRAVELING	*1.47*		*TRAVELING*	*18.4*
ECONOMY	*8.82*		*ECONOMY*	*5.83*

Key phrases K_0			Key phrases K_1	
Norway	*100.0*		*Norway*	*100.0*
Trondheim	*87.5*		*Trondheim*	*79.6*
Sør-Trøndelag	*85.2*		*Sør-Trøndelag*	*78.4*
for	*39.8*		*politiet*	*24.1*
Nidelva	*38.6*		*for*	*22.2*
Espen Sandmo	*38.6*		*Nidelva*	*21.0*
Morten Karlsen	*38.6*		*Espen Sandmo*	*21.0*
politiet	*38.6*		*Morten Karlsen*	*21.0*
NTNU	*12.5*		*Borussia Dortmund*	*13.0*
Rogaland	*10.2*	→	*lionel messi*	*13.0*
nettportalen	*9.09*		*vif*	*13.0*
Stavanger	*9.09*		*robin van persie*	*13.0*
Internasjonalt hus	*7.95*		*bayern munchen*	*13.0*
Internasjonalt kontor	*7.95*		*lambert*	*13.0*
University of Stavanger	*7.95*		*mourinho*	*13.0*
S O Bragstads Plass	*7.95*		*ole gunnar solskjær*	*13.0*
United States	*6.82*		*arve lote*	*13.0*
mms	*6.82*		*manchester united*	*13.0*
sms	*6.82*		*dortmundhelten*	*12.3*
john stene	*6.82*		*eva hongshagen*	*12.3*
...	...		...	...
...	....		...	...

Figure 5: Old and new user profile for a particular user

The news articles recommended at time t_1 reflect the changes of the user profile and are now totally dominated by sports news. As shown on the right hand side of Table 3, the top 10 stories recommended are in fact sports news, of which three are about local affairs, three concern national sports and four relate to sports events outside Norway.

Because the user profiles are updated on the basis of the users' current behavior, they tend to improve over time as the system learns more about the users' interests and preferences. This learning effect is important, as it means that the news recommendations will also improve if the user spends more time using the app. The dynamic nature of news streams, however, poses some particular problems to the news recommender systems. The selection of news stories changes continuously, since old stories grow outdated and new stories are added as events unfold. This means that the set of recommended stories will change from one point in time to another, even if the user profile is not changed, simply because the set of available stories is not the same. Consequently, the success of a particular user profile is not only decided by the content of the profile, but also by the availability of articles consistent with the profile. If there are no desirable articles available, the user risks watering down her profile when quickly inspecting the not so interesting articles presented by the app.

Table 3: Top recommended articles change as user's profile is developing

	Top 10 news at t_0			**Top 10 news at t_1**	
1	*"Avslutter søk etter mann i Nidelva"*	*NEWS*		*"Dette blir ingen vanlig kamp for Rekdal"*	*SPORTS*
2	*"Søker etter mann i Nidelva"*	*NEWS*		*"Mancini sparket ett år etter ligagullet med City"*	*SPORTS*
3	*"Liverpool-legende i Namsos"*	*SPORTS*		*"Hvert mål koster eliteserieklubbene over en million kroner"*	*SPORTS*
4	*"Her er kong Haakons fluktbil"*	*NEWS*		*"Wigan rykker ned - Arsenal nærmer seg mesterligaen"*	*SPORTS*
5	*"Lettere skadd etter utforkjøring"*	*NEWS*		*"Dekket over tabber med å skjelle ut journalister"*	*SPORTS*
6	*"Posten må leie inn lastebiler"*	*NEWS*		*"Rooney buet ut av Uniteds parademarsj"*	*SPORTS*
7	*"Flere 5-åringer har hull i tennene"*	*NEWS*		*"Elisabeths drømmemål går verden rundt"*	*SPORTS*
8	*"Hittil har vi vært heldige"*	*NEWS*		*"Nå håper Aalesunds superspiss på mer"* TS)	*SPORTS*
9	*"Trondheim ble årets kulturkommune"*	*LIFESTYLE*		*"Stegavik til Kostad - om dama blir i Byåsen"*	*SPORTS*
10	*"Nå lever tjuvfiskerne farlig"*	*SPORTS*		*"Høgmo blir trener i Djurgården"*	*SPORTS*

4. Discussion

The experiments so far suggest that user profiles built from the content of user's already read news articles produce recommendations that are consistent with her general preferences. There are, however, a number of issues that affect the quality of the profiles as well as the quality of the recommendations in the next round.

Since the whole profiling process starts with the categories and key phrases associated with each news article, the quality of these categories and key phrases is of paramount importance. A simple k-nearest approach is used to classify the documents, and the tests so far show that annotated categories are very reliable. For the key phrases the situation is more challenging, as it is notoriously difficult to extract named entities and prominent phrases from domains that are characterized by rapidly changing terminologies and collapsed grammatical constructions. The evaluation of the initial Named Entity Recognition component shows an accuracy that is clearly not satisfactory, and we are now in the process of implementing an improved NER component.

The user profile construction process itself is complex with several strategies that need to be carefully calibrated and may also possibly interfere with each other:

- *Weighting scheme for user actions.* So far we have assumed that all user actions carry the same weight. Sensitivity tests indicate, however, that the *Opened article view* action dominates all other actions, and the other actions mostly serve to amplify the contribution from the *Opened article view* action. The one notable exception to this is the preview time action which added information about articles that were somewhat interesting to the user, but not interesting enough for her to access the full text. In practice, this additional information turned out to be substantial and amounts to 92% of the size of the final user profile.
- *Update frequency.* In the current system we do not have a clear policy for when a user profile is updated. We upload the existing profile when the user enters the app, and we update the profile when her session is finished. Since this implies that her latest articles will not normally be reflected in her profile, we are implementing a feature that allows the user to manually enforce an immediate update of her profile based on her ongoing session.
- *Merging old and new user profile.* Our formula for merging old and new profile vectors is rather coarse-grained using a simple multiplication constant between 0 and 1 for degrading the content of the old user profile. An alternative method would be to update each individual element of the old vectors on the basis of the time that has passed since its value was last set. This seems conceptually more correct, though it implies a computationally more expensive solution.
- *Short-term vs long-term user profiles.* Earlier research by Billsus and Pazzani (1999) and Liu, Dolan and Pedersen (2010) argues that user profiles need to be divided into short-term profiles and long-term profiles. Short-term profiles relate to popular topics such as big or surprising events and need to be updated rapidly as the events take place. Long-term interests account for the user's general preferences and are assumed to be fairly stable from one session to another. We have in our work only defined one user

profile that addresses mostly long-term interests, assuming that separate recommendation strategies based on freshness and collaborative filtering will bring in news that cater for the user's short-term preferences.

It is still early to conclude about the quality of the user profiles, as their influence on the final recommendations is difficult to separate from other aspects of the recommendation engine itself. In a hybrid recommenddation system there are several recommendation strategies that need to be weighted and combined to produce the final results (Borges and Lorena, 2010). In our case, the weights of freshness, geographical proximity, collaborative filtering and content-based filtering will severely affect the impact of the learned user profiles, and any changes to these weights may necessitate adjustments to how these profiles are generated.

A separate issue is the relative weighting of categories and content in the user profile, which also affects the profile's effect on the recommendations made. Currently these weights are assumed to be equal, though the weights may need to be further refined as part of an analysis of the recommendation system's total weighting scheme.

Finally, there are two parties involved in news recommendation, the news *provider* and the news *reader*, that evaluate the quality of recommendations from two very different perspectives. Whereas the news reader is mostly interested in getting only news consistent with her current interests, the news provider wants to present news that extend her interests and thereby hold on to them as active news app users. Serendipity is important for this reason, but also the recommendation of not so relevant news in cases where no new relevant stories have come in.

Ultimately, news providers measure the success in terms of click-through rates, even though these may only be moderately correlated with the readers' perception of news relevance.

5. Conclusion

We have in this paper presented an approach for learning user profiles from observing the user's own actions on a mobile news app. No explicit information about user preferences is given or retrieved from other sources in this process. The user profiles are afterwards used by a hybrid news recommender engine to produce a personalized mobile news service.

Current research on recommendation technologies has had an emphasis on recommendation strategies and there is only limited research on the construction of user profiles from mobile user behavior. Our approach analyzes the content of news articles and associates the users with user profile vectors that aggregate the contents of previously read articles. The two profile vectors, one for representing category preferences and another for representing content preferences, are both extracted using standard text mining techniques for classification and entity extraction.

The results so far suggest that the user profile captures important aspects of the user and leads to recommendations more consistent with her general preferences. These profiles, however, typically support the content-based recommendation part of the news app, and there are other strategies that should deal with short-term issues and news relevant to the geographical neighborhood. The weighting of recommendation strategies in such a hybrid recommender system is challenging and a topic for further research.

In our further work we plan to refine the construction of user profile vectors and evaluate different strategies for user action weighting, update frequencies, and merging of profile vectors. The same user profiles will also gradually be extended to support collaborative filtering, which means that they may need to incorporate aspects that have so far not been needed for content-based filtering.

Acknowledgements

This research was supported by Telenor Group as part of their collaboration with the Department of Computer and Information Science at the Norwegian University of Science and Technology in Trondheim, Norway.

References

Adomavicius, G. and Tuzhilin, A., 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), pp. 734-749

Belkin, N. J. and Croft, W. B., 1992. Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM,* 35(12), pp. 29-38

Billsus, D. and Pazzani, M. J., 1999. A hybrid user model for news story classification. In *Proceedings of the Seventh International Conference on User Modeling* (UM'99). Berlin-Heidelberg: Springer. pp. 99-108

Billsus, D. and Pazzani, M. J., 2000. User Modeling for Adaptive News Access. *User Modeling and User-Adapted Interaction*, 10, pp. 147-180

Borges, H.L. and Lorena, A. C., 2010. A Survey of Recommender Systems for News Data. In: Szczerbicki. E., ed. *Smart Information and Knowledge Management*, SCI 260. Berlin-Heidelberg: Springer. pp. 129-151

Brasethvik, T. and Gulla, J. A., 2002. A conceptual modeling approach to semantic document retrieval. In: *Proceedings of the 14th International Conference on Advanced Information Systems Engineering (CAiSE'02).* Berlin-Heidelberg: Springer. pp. 167-182

Cantador, I., Bellogin, A. and Castells, P., 2008. Ontology-Based Personalised and Context-Aware Recommendations of News Items. In: *Proceedings of the 7th International Conference on Web Intelligence.* IEEE. pp. 562-565

Circa, 2013. *Catch up quick.* [online] Available at: http://cir.ca/. [Accessed 18 April 2013]

Das, A. S., Datar, M., Garg, A. and Rajaram, S., 2007. Google news personalization: scalable online collaborative filtering. In: *Proceedings of the 16th international conference on World Wide Web.* ACM. pp. 271-280

De Francisci Morales, G., Gionis, A. and Lucchese, C., 2012. From chatter to headlines: harnessing the real-time web for personalized news recommendations. In: *Proceedings of the Fifth ACM international conference on Web search and data mining.* ACM. pp. 153-162

Gulla, J. A., Auran, P. G. and Risvik, K. M., 2002. Linguistic Techniques in Large-Scale Search Engines. In: *Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems (NLDB'02)*, pp. 218-222

Haugen, K. R., 2013. *Mobile News: Design, User Experience and Recommendation.* MSc thesis. Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim

Jannach, D., Zanker, M., Felfernig, A. and Friedrich, G., 2010. *Recommender Systems: An Introduction.* Cambridge University Press

Kim, H. R. and Chan, P. K., 2003. Learning implicit user interest hierarchy for context in personalization. In: *Proceedings of the 8th international conference on Intelligent user interfaces.* ACM. pp. 101-108

Lee, U., Liu, Z. and Cho, J., 2005. Automatic identification of user goals in web search. In: *Proceedings of the 14th international conference on World Wide Web.* ACM. pp. 391-400

Liu, J., Dolan, P. and Pedersen, E.R., 2010. Personalized news recommendation based on click behavior. In: *Proceedings of the 15th international conference on intelligent user interfaces.* ACM. pp. 31-40

Nasraoui, O., Soliman, M., Saka, E., Badia, A. and Germain, R., 2008. A web usage mining framework for mining evolving user profiles in dynamic web sites. *IEEE Transactions on Knowledge and Data Engineering.* 20(2), pp. 202-215

Nilsen, J.E.B., 2013. *Large-Scale User Click Analysis in News Recommendation.* MSc. Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim

O'Banion, S., Birnbaum, L. and Hammond, K., 2012. Social media-driven news personalization. In: *Proceedings of the 4th ACM RecSys workshop on Recommender systems and the social web.* ACM. pp. 45-52

Rajaraman, A. and Ullman, J. D., 2011. *Mining of Massive Datasets.* Cambridge University Press

Schafer, J. B., Konstand, J. and Riedl, J., 1999. Recommender Systems in E-Commerce. In: *Proceedings of the 1st ACM conference on Electronic Commerce (EC'99).* New York: ACM. pp. 158-166

Shuai, X., Liu, X. and Bollen, J., 2012. Improving news ranking by community tweets. In: *Proceedings of the 21st international conference companion on World Wide Web.* ACM. pp. 1227-1232

Solskinnsbakk, G. and Gulla, J. A., 2010. Combining ontological profiles with context in information retrieval. *Data & Knowledge Engineering*, 69(3), pp. 251-260

Speretta, M. and Gauch, S., 2005. Personalized search based on user search histories. In: *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence.* IEEE. pp. 622-628

Tavakolifard, M., Gulla, J. A., Almeroth, K. C., Ingvaldsen, J. E., Nygreen, G. and Berg, E., 2012. Tailored News in the Palm of your HAND: A Multi-Perspective Transparent Approach to News Recommendation. In: *Proceedings of 22nd International World Wide Web Conference (WWW'13), Companion Volume.* Rio de Janeiro. pp. 305-308

Thurman, N., 2011. Making 'The Daily Me': Technology, Economics and Habit in the Mainstream Assimilation of Personalized News. *Journalism: Theory, Practice & Criticism*, 12(4), pp. 395-415

Yeung, K., 2013. *News curator Summly launches to help simplify the way we consume news on mobile devices.* [online] Available at: http://thenextweb.com/apps/2012/11/01/news-curator-summly-launches-to-help-simplify-the-way-we-consume-news-on-mobile-devices/#!p1eeM. [Accessed 17 April 2013]

JPMTR 026 | 1314
UDC 004.42:004.58

Original scientific paper
Received: 2013-08-01
Accepted: 2013-10-25

UPCV - Distributed recommendation system based on token exchange

Ville Ollikainen[1], *Aino Mensonen*[1], *Mozhgan Tavakolifard*[2]

[1] VTT Technical Research Centre of Finland
P. O. Box 1000
FIN-02044 VTT, Espoo, Finland

E-mails: ville.ollikainen@vtt.fi
aino.mensonen@vtt.fi

[2] Department of Computer and Information Science
Norwegian University of Science and Technology
Sem Sælands vei 7-9
N-7491 Trondheim, Norway

E-mail: mozhgan@idi.ntnu.no

Abstract

Most conventional recommendation systems are based on service-specific data repositories containing both user and item data. In this paper, we introduce an alternative approach called UPCV (Ubiquitous Personal Context Vectors) that inherently supports distributed computing and distributed data repositories. The principal idea is that each user-item interaction can update the data associated with *both* the user *and* the item. When updating, item data is made to slightly resemble user data and vice versa, leading to increasing similarity between them. Through interactions, similarity will spread from users to items, from items to users, making it possible to inherently provide user-item, item-item, item-user and user-user recommendations. The principle introduced in this paper can be used as a baseline for the design of different types of collaborative recommender systems. The main advantages of this method are that it requires no content analysis, preserves users' privacy and supports scalability. The method was evaluated using data from 1 575 book club members: the members were asked which books they had read and liked. The quantitative analysis indicates that the most promising results are obtained for active readers. However, even for less active readers and without content analysis, the recommendation list tends to be populated by the same authors and/or authors of the same genre that the readers have liked, leading to meaningful recommendations.

Keywords: recommendation, collaborative filtering, distributed computing, cloud computing, scalability, privacy, deniability

1. Introduction and background

Recommendations have become an integral part of successful web services. Recommendation systems are used on one hand for e-commerce (e.g., Amazon) or advertising (e.g., Google) and on the other hand to improve user experience (e.g., Netflix). The more the amount of information in a service increases, the more important it becomes to help users discover what is most relevant for them.

The recommendation problem can be defined as estimating a user's response to new items based on historical information stored in the system, and suggesting novel and original items for which the predicted response for that particular user is high (Desrosiers and Karypis, 2011). Prediction of user interests is traditionnally through demographic data, such as age, sex, income level and matrimonial status. The availability of more data has led to more sophisticated recommending algorithms being proposed in the literature, most commonly classified into two basic categories: content-based and collaborative recommendations. Content-based recommenders are based on representing the items with a set of attributes and using these attributes to find the most relevant content for a particular user. Collaborative recommendations, on the other hand, learn from the behaviour of users. Some of the more recent novel recommendation techniques use data from social networking (Golbeck, 2006; Liu and Lee, 2010) or use hybrid models merging several techniques (Bobadilla et al., 2013).

This work presents a novel method, UPCV (Ubiquitous Personal Context Vectors). The method models each user and item with a set of tokens, each token carrying a random value. Interaction between user and item results in randomly selected tokens being copied from the token set of the user to the token set of the item, and vice versa. Each interaction increases the number of common tokens among these token sets. When the same user interacts with several items, or the same item is involved in interactions with several users, these common token numbers are spread around, resulting in closer similarities among different token sets in the system. Since tokens spread in interactions, it is likely that similarities between two token sets originate from similar user behaviour. No content analysis is required.

Despite significant advances in the field of contemporary recommender systems, there still remain challenges that limit the effectiveness of these systems. Our proposed model targets these challenges by providing:

- **A broader view of user behaviour:**
  User data gathered from a single service has only a narrow coverage of user behaviour.

- **Domain knowledge independency:**
  Some approaches (e.g., Bäck, 2010) suggest storing personal profiles in a database and delivering from there in order to authorize parties providing personalized services. As such, personal profiles support content-based recommenders, matching user interests to what is available in the service. Content-based approaches require knowledge of the domain in order to match user and item data efficiently. Such techniques have a natural limit in the number and type of features associated, whether automatically or manually, with the objects they recommend. There is a frequent need for domain knowledge (of actors and directors in movie recommendations, for example) and occasionally for domain ontologies (Lops et al., 2011).

- **Preserving user privacy:**
  Privacy concerns have been raised both by recommendation systems gathering data from several services (such as Apple IFA; Stampler, 2012) and by recommendations running on social networking sites.

- **Distributed and cloud based computing:**
  In general, recommender systems are based on service-specific data repositories containing both user and item data. Despite recent development in distributed and cloud computing, single repositories pose an inherent problem in terms of scalability.

Moreover, we here report on an evaluation of UPCV based on collecting data from 1575 book club members. The quantitative results indicate that the most promising recommendations are obtained for active readers: readers with more than 28 book selections in the training data would have expected over 50 % probability of obtaining a successful recommendation in a list containing no more than five books.

However, even for less active readers and with no content analysis, the recommendation list tended to be populated by the same authors and/or authors of the same genre that they had liked, leading to meaningful recommendations. The remainder of this paper is structured as follows: The novel recommendation method based on data fusion is described in section 3 and evaluated in section 4. Discussion and conclusions are in sections 5 and 6, respectively.

2. Related work on item based collaborative filtering method

In this section, we analyse different item based recommendation generation algorithms. We look into different techniques for computing item-item similarities (e.g., item-item correlation vs. cosine similarities between item vectors) and different techniques for obtaining recommendations from them (e.g., weighted sum vs. Regression model).

The item based approach (Sarwar et al., 2001; 2002; Su and Khoshgoftaar, 2009; Linden et al., 2003; Miyahara and Pazzani, 2002; O'Connor and Herlocker, 1999; Xue et al., 2005; Deerwester et al., 1990) looks into the set of items the target user has rated and computes how similar they are to the target item, thereafter selecting the most similar items. At the same time their corresponding similarities are also computed.

Once the most similar items are found, the prediction is then computed by taking a weighted average of the target user's ratings on these similar items. The basic idea in similarity computation between two items *i* and *j* is to first isolate the users who have rated both of these items and then to apply a similarity computation technique to determine the degree of similarity. The most popular methods for calculating the similarity are: cosine-based similarity, correlation-based similarity and adjusted cosine similarity. In the following, we briefly describe each similarity method (Sarwar et al., 2001):

- Cosine-based similarity: Two items are represented by two vectors in the m-dimensional user space. The similarity between them is measured by computing the cosine of the angle between these two vectors.

- Correlation-based similarity: Similarity between two items *i* and *j* is measured by computing the Pearson-r correlation. To make the correlation computation accurate, we must first isolate the co-rated cases (i.e., cases where the users have rated both).

- Adjusted cosine similarity: Computing similarity by using a basic cosine measure in an item based case has one important drawback: the differences in (rating) scale between different users are not taken into account. The adjusted cosine similarity addresses this drawback by subtracting the corresponding user average from each co-rated pair.

Since the item based approach requires at least one user having rated both items *i* and *j*, the computation is possible only for a limited set, leading to limited coverage which is a common problem in collaborative filtering methods, addressed by e.g., Choi and Suh (2013) and Desrosiers and Karypis (2011).

Once the set of most similar items is isolated (based on the similarity measures), the next step is to look into the target user ratings and use a technique to obtain predicttions.

Two popular approaches are Weighted sum and Regression as explained below (Sarwar et al., 2001):

- Weighted sum: The prediction of an item for a user is computed as the sum of the ratings given by the target user on the similar items. Each rating is weighted by the corresponding similarity between the items. This approach tries to capture how the target user rates the similar items. The weighted sum is scaled by the sum of the similarity terms to make sure the prediction is within the predefined range.
- Regression: The basic idea here is to use the same formula as the weighted sum technique but, instead of using the similar item rating values, this model uses their approximated values based on a linear regression model.

  In practice, the similarities computed using cosine or correlation measures may be misleading in the sense that two rating vectors may be distant (in Euclidean sense), but may yet have very high similarity. In such a case, using the raw ratings of the "so-called" similar item may result in poor prediction.

In summary, the main advantages of item based collaborative filtering methods are that there is no need to consider the content of the items being recommended and that these approaches scale well with co-rated items.

In general, the main shortcomings of these methods are the lack of ability to make recommendations for new users and new items, and the limited scalability for large datasets (Su and Khoshgoftaar, 2009).

3. The proposed method

We now introduce a novel recommendation method based on data fusion, UPCV, most closely related to memory based collaborative filtering in the sense defined in the comprehensive survey of recommender systems by Bobadilla et al. (2013). As such, our approach requires no content analysis.

In UPCV, each user and each item (e.g., a news article) is associated with a set *A* of tokens *x* (Equation 1), each represented as a 32-bit integer ($x \in \mathbb{Z}$).

$$A = \{x \in \mathbb{Z} \mid < 0 \le x < 2^{32}\} \qquad [1]$$

When new users or new items appear, their token sets are initialized to contain one single random value. In an interaction between a user *i* and an item *j* (e.g., when user *i* reads the news article *j*), a small number of tokens X_i are *copied* from the token set of user A_i to the token set of item A_j, see Equation 2.

The reverse also happens, see Equation 3.

$$A_j' = A_j \cup X_i,\ X_i \subseteq A_i \wedge |X_i| \le {}^{15}/_{100}\ Max\{|A_j|\} \qquad [2]$$

$$A_i' = A_i \cup X_j,\ X_j \subseteq A_j \wedge |X_j| \le {}^{15}/_{100}\ Max\{|A_i|\} \qquad [3]$$

We define the maximum cardinality of a token set as 256. Prior to reaching this limit, randomly selected tokens from the receiving token set are deleted so as not to exceed the maximum number. We furthermore limit the number of copied tokens $|X|$ to 15 % of the maximum size of the receiving token set. The selection of this percentage does not appear to be critical since some of our experiments were made using 5 % and 10 % limits, leading to very similar results.

The result of this procedure is an increasing number of common tokens in the respective token sets after each interaction. When the same user interacts with several items, or when the same item is involved in interactions with several users, the token numbers are spread around, resulting in similarities among different token sets in the system. Since tokens spread in interactions, *it is likely that similarities between two token sets originate from similar user behaviour*. Since there is no limit on how far tokens may spread, recommendation coverage is not limited.

The similarity $S(A_i, A_j)$ between the token sets of user *i* and item *j* $S(A_i, A_j)$ is measured using the Jaccard similarity measure (Equation 4).

$$S(Ai, Aj) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|} \qquad [4]$$

From this point on, the making of recommendations is straightforward, no matter whether they are user-to-user (finding users of similar behaviour), user-to-item (finding items that may interest the user), item-to-item (finding items of similar interest) or item-to-user (finding users who might be interested in an item); it is only necessary to find the token sets (from the total population of size *n*) that are most similar to the given token set. Thus the recommended item *j* for a given token set A_i is defined by Equation 5.

$$j = \underset{k=1..n, i \neq k}{arg\,max}\ S(A_i, A_k) \qquad [5]$$

In contrast to traditional recommenders, our method inherently supports distributed computing and distributed data repositories. Figure 1 illustrates a concept for storing user data (token set A_i) at the terminal end, while item data (token set A_j) resides at the server end.

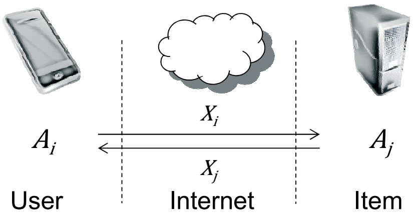


Figure 1: Conceptual illustration of a data transaction between user and item data

Tokens exchanged in a transaction (X_i and X_j respectively) are transferred over the internet. The tokens are fundamentally random numbers and irreversible, they carry no history, thus exchanging the data poses no threat to privacy. This arrangement is motivated by the notion that both parties - the user or the owner of the item(s) - own and have control over their own recommendation data. We consider this a substantial advantage over the mainstream recommendation systems, since it is not only users who may be aware of their privacy, but also enterprises which are reluctant to disclose any business critical information to third parties, such as to an ecosystem owner.

Figure 2 illustrates a concept for obtaining recommendations from a web service. In this case the user submits his/her token set A_i to a web service that has the potential to provide several items (token sets A_j respectively). Such items are recommended for the user which have token sets with the smallest distance to the token set provided.

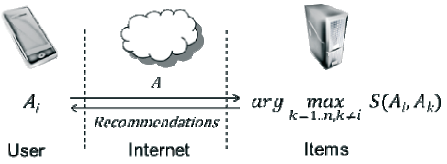


Figure 2: Conceptual illustration for obtaining a recommendation

Once again, all recommendation data for the items in the depicted web service is owned by the web service itself. From the user point of view, the token set A_i is a *de facto* representation of the user profile in an abstract format, basically inherited from other users with somewhat similar behaviour. Therefore, no definite conclusions of personal preferences or history can be drawn.

Furthermore, even with the absence of the user token set A_i, such as in a case of a new user, the depicted web service may generate item-to-item ("see also") recommendations by searching for the smallest distances from the token set of the item *j* (Equation 6).

$$\hat{j} = \underset{k=1..n, j \neq k}{arg\,max}\, S(A_j, A_k) \qquad [6]$$

Since the computation involves only the user terminal and the server in question, the architecture can be considered to inherently comply with internet services and scale up respectively.

4. Evaluation

4. 1 Data sets

For evaluation of the UPCV method we used data from 1 575 book club members. The members were asked which books they had read and liked. We aimed to predict the books (hereinafter "items") a user might be interested in by hiding part of the questionnaire data from the recommender for use in validation only. We therefore divided the sparse data randomly into training and validation data sets. Based on the training set, we generated recommendations for each member aiming to predict which books the member would have in the validation data set.

4. 2 Data gathering

We arranged an online survey about favourite books. Bonnier Books Finland provided us with a list of 1 041 books that have been available for their book club members. This list was divided into a set of shorter lists, based on author names, A-D first, E-H next, etc. The questionnaire asked respondents to select books they had "read and liked".

They were able to select as many books from as many lists as they wished. A link to the online questionnaire was sent to the book club members. 1 575 book club members responded to the questionnaire. The total number of individual selections was 55 434, leading to an average of 27.6 selections per respondent. The standard deviation was 25.9, indicating that we had both active and inactive readers among the respondents.

Figure 3 illustrates the distribution.

The selections were converted into user-item pairs, shuffled into random order and divided into two groups, each consisting of 27 717 user-item pairs in random or-

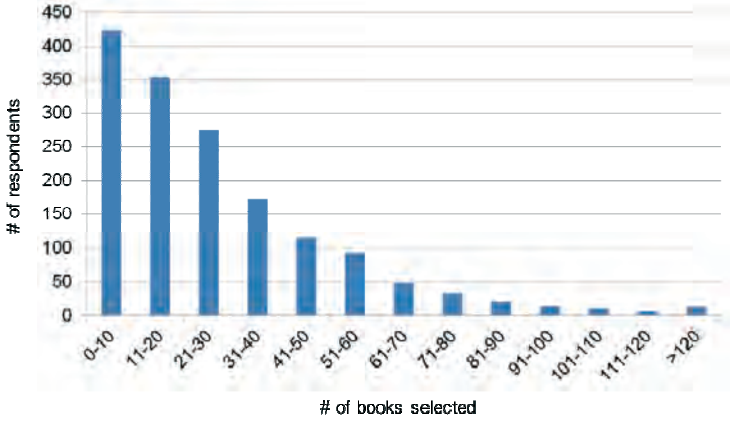


Figure 3: Distribution of the number of books (items) selected by respondents (users)

der. We used the first group as training data, while the second group remained for validation. 1 481 users had at least one selection in both training and validation data.

Token sets for all users and all items were created by entering the user-item pairs of the training data in the recommendation system as an interaction between user and item. We then generated recommendations for each user by searching token sets of the items that had the smallest distance to the token set of the user.

4.3 Quantitative analysis

A recommendation list of length N is considered successful if there is a match between a respective user-item pair in the test data and the N'th recommendation for the user on the list. The search starts from the beginning of the list, on which highest recommendation with smallest distance are provided first.

Figure 4 illustrates that recommendation lists were shorter for users who had a higher number of books selected in the training data. For example, provision of a recommendation list with 5 books would have been achieved with 50 % probability when a user had 28 books in the training data.

Each dot in the figure represents a group of at least 30 users. The figure also illustrates intervals that exclude the upper and lower 10 % of the users in each group, and a best matching trend line (power type).

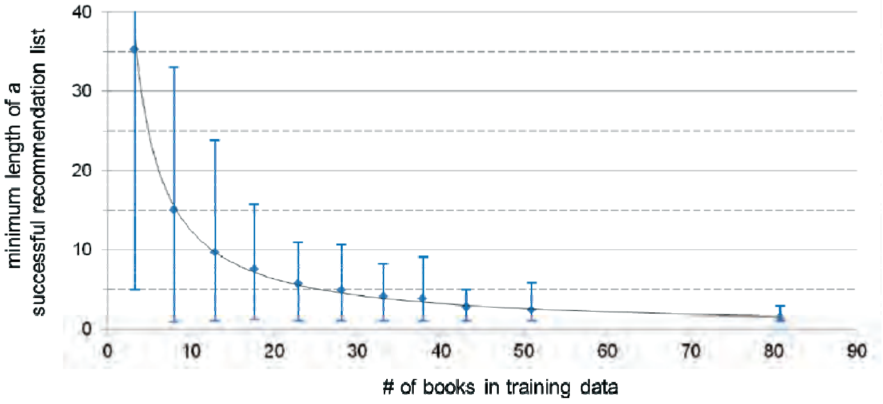
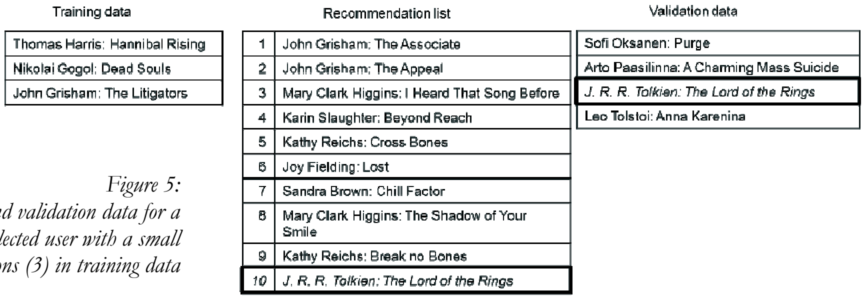


Figure 4: Length of a successful recommendation list (vertical axis) when a user had a certain number of books selected in the training data (horizontal axis)

4.4 Qualitative observations

Regarding recommendations made for users with only a small number of selections in the training data, the quantitative results indicate that a rather long list would be necessary for proper recommendation. However, even in these cases it seems that the same authors and/or authors of the same genre they liked has a tendency to populate the recommendation list. Figure 5 illustrates one randomly selected example.

Although the quantitative length of a successful recommendation is 10, most books in the recommendation list belong to the same genres (thrillers, crime fiction) as two out of the three books in the training data.



Training data

Training data
Thomas Harris: Hannibal Rising
Nikolai Gogol: Dead Souls
John Grisham: The Litigators

Recommendation list

	Recommendation list
1	John Grisham: The Associate
2	John Grisham: The Appeal
3	Mary Clark Higgins: I Heard That Song Before
4	Karin Slaughter: Beyond Reach
5	Kathy Reichs: Cross Bones
6	Joy Fielding: Lost
7	Sandra Brown: Chill Factor
8	Mary Clark Higgins: The Shadow of Your Smile
9	Kathy Reichs: Break no Bones
10	*J. R. R. Tolkien: The Lord of the Rings*

Validation data

Validation data
Sofi Oksanen: Purge
Arto Paasilinna: A Charming Mass Suicide
J. R. R. Tolkien: The Lord of the Rings
Leo Tolstoi: Anna Karenina

Figure 5: Training and validation data for a randomly selected user with a small number of selections (3) in training data

5. Discussion

We have described a simple collaborative recommender method based on token exchange and designed to protect privacy and support scalability in a distributed architecture. We have evaluated its performance by applying it to questionnaire data from a book club survey.

The quantitative results were better for the most active users. Active readers with more than 28 book selections in the training data had over 50 percent probability of obtaining a successful recommendation in a list containing no more than five books.

Even for less active readers, the recommender system provided selections of a similar genre based on the training data related to the user. We emphasize that these result were obtained *without content analysis*. The only data a book received initially was a single token containning one random integer.

The results indicate that the tokens successfully distributed by interactions among users and items, together with comparison of various token sets, can provide meaningful insights for recommendations.

6. Conclusions and future work

Our study provides an overview of and evaluation results for our proposed approach, which is based on exchanging tokens. Use of this method can easily be extended to other application areas since it is not dependent on any particular assumption about the application area. Exchanging tokens in social networking sites, for instance, might lead to similar tokens for people in the closest social network and - consequently - a higher probability of their receiving similar recommendations.

The method is efficient enough to learn from fairly few interactions, as described in the previous example of a reader with only a couple of books in the training data. With its short required learning time, the method can be beneficial for temporary content, such as news articles.

Generally, there seems to be a trade-off between privacy, trust and recommendation quality. If a service has a comprehensive view of the behaviour and preferences of the user, recommendations may become very accurate. However, in this case, the user must have an indisputable trust in the service, otherwise privacy is compromised. Our approach has an inherent advantage in this respect, since registering any history of actual interactions is not required. Therefore, as a last resort for privacy, the tokens may be said to have been inherited from any interaction; our approach provides a fair degree of deniability.

Further studies are necessary to investigate how to assign tokens to item properties (e.g., keywords of a news article, or its semantic network). Rather than an interaction with the article itself, reading would initiate a series of interactions with various properties. Recommendations would then aggregate results by finding items with the most highly ranked properties for the user.

Moreover, we aim to use considerably larger data for evaluation. It is likely that related data mining studies (e.g., Segond and Borgelt, 2011) may further help in improving and optimizing UPCV.

Acknowledgements

In addition to the authors, a number of people have made the study possible: thus, we thank Minna Kokka (Bonnier Books Finland) and Katri Grenman (VTT) for arranging the survey and our colleagues at VTT, especially Raimo Launonen, Caj Södergård and Atte Kortekangas, for contributing to the development of UPCV, and Bidyut Patra and Alain Boyer for most valuable comments in finalizing this paper.

We also thank VTT's Open Smart Spaces programme (especially Heikki Ailisto), the Tekes Ubicom programme (especially Pekka Yrjölä), the Digile Oy research programme Next Media (especially Eskoensio Pipatti), and the ENPI project Digital Sphere - all these projects and the respective persons for making the UPCV development possible.

References

Bobadilla, J., Ortega, F., Hernando, A. and Gutiérrez, A., 2013. Recommender systems survey. *Knowledge-Based Systems*, 46(0), pp. 109-132

Bäck, A., 2010. Käyttäjän itsensä hallitsema semanttinen profiili verkkopalveluiden personointiin. *GT-lehti*, No. 4, 2010. Espoo: VTT, pp. 25-27

Choi, K. and Suh, Y., 2013. A new similarity function for selecting neighbors for each target item in collaborative filtering. *Knowledge-Based Systems*, 37(0), pp. 146-153

O'Connor, M. and Herlocker, J., 1999. Clustering items for collaborative filtering. *Proceedings of the ACM SIGIR Workshop on Recommender Systems.* UC Berkeley, pp. 121-128

Deerwester, S. T., Dumais, S., Landauer, T. K., Furnas, G. W. and Harshman, R., 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), pp. 391-407

Desrosiers, C. and Karypis, G., 2011. *A Comprehensive Survey of Neighborhood-based Recommendation Methods.* Boston, MA: Springer US

Golbeck, J., 2006. *Generating predictive movie recommendations from trust in social networks.* Berlin: Springer-Verlag

Linden, G., Smith, B. and York, J., 2003. Amazon.com recommendation - Item-to-item collaborative filtering *IEEE Internet Computing*, 7(1), pp. 76-80

Liu, F. and Lee, H. J., 2010. Use of social network information to enhance collaborative filtering performance. *Expert Systems with Applications*, 37(7), pp. 4772-4778

Lops, P., De Gemmis, M. and Semeraro, G., 2011. *Content-based Recommender Systems: State of the Art and Trends.* Boston, MA: Springer US

Miyahara, K. and Pazzani, M. J., 2002. Improvement of Collaborative Filtering with the Simple Bayesian Classifier Information Processing Society of Japan. *IPSJ Journal*, 43(11), pp. 3429-3437

Segond, M. and Borgelt, C., 2011. *Item Set Mining Based on Cover Similarity.* Berlin, Heidelberg: Springer

Stampler, L. 2012. Here's Everything We Know about IFA, The iPhone Tracking Technology in Apple's iOS 6. *Business Insider*, October 15, 2012. [online] Available at http://www.businessinsider.com/everything-we-know-about-ifa-and-tracking-in-apples-ios-6-2012-10 [Accessed 23 July 2013]

Sarwar, B., Karypis, K., Konstan, J. and Riedl, J., 2001. Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th international conference on World Wide Web*, pp. 285-295

Sarwar, B., Karypis, K., Konstan, J. and Riedl, J., 2002. Recommender Systems for Large-Scale E-Commerce: Scalable Neighborhood Formation Using Clustering. *Proceedings of the Fifth International Conference on Computer and Information Technology*

Su, X. and Khoshgoftaar, T. M., 2009. A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence*, vol. 2009, Article ID 421425. doi:10.1155/2009/421425

Xue, G.-R., Lin, C., Yang, Q., Xi, W., Zeng, H.-J., Yu, Y. and Chen, Z., 2005. Scalable Collaborative Filtering Using Cluster-based Smoothing. *Proceedings of the ACM SIGIR Conference*, pp. 114-121

Topicalities

Edited by Raša Urbas

Contents

News & more

Braille application coating

Jowat AG has developed new coating product for special "Braille" applications - Perma Coat®. This innovative UV-curing system complies with present standards and superior processing parameters. Printed surfaces with Braille code can be manufactured by an innovative method - differing from conventional high-relief embossing - coating. The technology allows application of Braille letters as coating dots using a nozzle applicator, which simplifies the production processes while still complying with the respective standards.

The new Jowat product ensures meeting of all current demands of valid standards for high quality performance with a Braille dot heights of more than 0.5 mm. the wide adhesion of this coating, in combination with ist excellent flow properties through the nozzle and a fast hardening performance under UV light, facilitate inline processing and allow high line speeds for the manufacture of folding boxes.

The Perma Coat® coating remains permanently transparent, thereby ensuring a perfect visual appearance of the folding box. The tactile effect of the lacquered Braille dots promotes a clear readability of the information on the package. Apart from the "folding box" applications this product could be used for any and all printed products which can be finished showing Braille letters.

New developments of organic and printed electronics

Plastics with modifiable material properties, dimensionally stable as thermoplastics, thermosets or elastomers, films or coatings, granular or expanded, are an indispensable part of everyday life. Plastics' structural diversity is now being augmented by a further dimension: with suitable molecular configuration, they can also be used as electrical conductors and semiconductors (albeit with still limited mobility of the charge carriers). They thus serve as system components of "organic" and "printed" electronics. "Organic" because their transistors, sensors and LEDS are not based on silicon or gallium arsenide, but on carbon derivatives. And "printed" because two-dimensional circuit patterns can be printed "from the reel" with structural fineness of just a few tens of micrometres onto flexible and also transparent substrates by using conventional mass printing processes (flexo, screen-printing, inkjet).

Integration in objects

This yields electronically or photonically functionalised surfaces, three-dimensionally on all conceivable objects including even textiles. They form capacitive touch sensors, large-area luminous fields with OLEDs (organic light-emitting diodes), sensors and detectors for environmentally or medically important data such as temperature and humidity. They operate as organic solar cells, or as flat, light printed batteries for a wide range of miniaturised devices.

Advanced wide format printer

Xerox has produced a new advanced large format inkjet printer for printing indoor posters, signs, point-of-purchase graphics and banners -IJP 2000. High speed is distinguished with stationary print heads which allow paper to move under five print heads in one single pass.

Faster printing speed offers printing of up to 420 square meters per hour. High-quality color signs can so be printed in five seconds, 9 meter banners in one minute and production runs of 200 prints in about 20 minutes. Fast - instantly dry ink enables production of variety of full-color products with crisp, precise imagery.

Certified eco inks

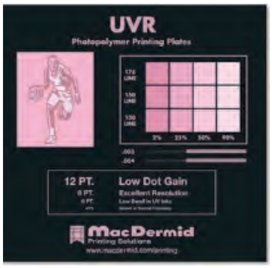
Mimaki's Sb300 series of dye sublimetion inks has been awarded with ECO-passport certification by Centexbel. This certification, which certifies compliance with the Oeko-tex standard 100 and ECO-passport certification system, assures users that inks cause no harm when contacted with skin.

Sb300 ink is aqueous dye sublimation ink which has been recently introduced to the market.

The ink is optimized for use with the Mimaki TS500-1800 textile printer for dye sublimation transfer. Mentioned printer can operate in a six color mode for smooth gradations and higher quality light colors, or a four color mode for increased productivity.

New digital flexo UVR printing plate

MacDermid has introduced a new digital flexo UVR printing plate designed specifically for use with UV inks. This plate is swell-resistant in aggressive UV inks and also blasts low dot gain and high resolution capabilities. The UVR 55 durometer plate has exceptional drape which allow the plate to wrap around a narrow web cylinder with smaller repeats with relative ease, reducing the risk of plate fix.



It is suited for the narrow web market, where UV inks are heavily used. UVR plate performance is also comparable in water-based inks and is compatible with MacDermid's LUX® Platemaking Process.

New plates are commercially available in thick-nesses of 1.14 and 1.7 mm and in format sizes up to 1.321 x 2.032 mm.

PUR product adhesives

Planatol has introduced three new polyurethane (PUR) graphic adhesives suitable for production - gluing at lower temperatures of standard graphic products, photo books, illustrated books and other high-quality books.

Planatol 2880 PUR is an adhesive for the back spine and it enables excellent adhesion even on difficult papers. Planatol 1142 PUR also offers strong adhesion while being soft enough to enable very good lay-flat behavior.

This adhesive helps reducing energy and emissions. It can be used without restrictions at working temperatures below 100 °C.

Application driving forces

The driving forces in the development of applications can be found in the automotive, pharmaceutical, consumer electronics and "smart" packages for foods, medicines and other consumer items. With inexpensive printed, radio-frequency identification (RFID) tags, smart packages are capable of making merchandise management more efficient and, with dynamically updated display fields, of informing the consumer of the best-before date, drawing attention to gaps in the cooling chain for sensitive goods and guaranteeing the authenticity of high-grade articles by establishing links to traceable supply chains.

Next in line are organic displays and touch sensors in premium class cars as replacements for mechanical indicators and switches. Then there are reversing lights with OLEDs, so that today's LED lights can be replaced to save energy.

Flexible displays for e-readers

Some e-readers with "electronic paper" from E-Ink enjoy widespread popularity because of the energy-efficient, bistable principle of their electrophoretic displays. They are essentially ideal for presenting static content such as book pages.

The next development step will bring forth lighter, flexible and maybe even roll-up e-readers and tablets without the heavy cover glass. The most progress here has been made by Plastic Logic that produces backplanes of organic thin film transistors (OTFTs), i.e. the active matrix for individual pixel control.

What is still hampering the development of organic photovoltaics and display technology is their hermetic encapsulation to provide protection from atmospheric water vapour that corrodes their electrodes and shortens their service life. The solution is laminated barrier films, for which transparent layers of amorphous silicon dioxide appear to be very well suited.

OLED lighting

OLED light sources are competing with established LEDs and halogen lamps. They promise dynamically colour-controllable light emitted uniformly over a large area and can be attached in architecturally attractive ways to the surfaces even of familiar objects in the home.

Organic photovoltaics and batteries

Organic photovoltaics (OPV) is already commercially available as local supply sources for mobile data and consumer devices. The long-term prospects include applications in the envelopes of vehicles and buildings (BIPV, building-integrated photovoltaics).

Available as system components are printed data memories - in the form of the ferroelectric, non-volatile memory films of the Norwegian manufacturer Thinfilm. These can be combined with a transistor logic produced at contract researcher PARC in California to yield software-addressable memory modules. With a printed thermistor as a temperature sensor and a display field together with a printed battery, they can be extended into a complete measuring system.

Printed batteries are also a focus of system integration. They can be integrated with display and luminous fields, touch sensors and solar cells in packages, textiles and other consumer items, elevating them to new levels of value and functionality.

Excerpts from a professional article published by Messe Düsseldorf

Color quality solution

QuadTech has in cooperation with Huber group developed a new solution - the new Color Quality Solution which offers faster and reliable method for ensuring color control.

Printing control performed directly in the printing system with in-line spectral data obtained with control cameras, ink formulation data from X-Rite and ink dispensing system.

QuadTech
Our passion is your success.
huber group

Color Quality Solution is designed on special HD quality control cameras of high resolution. Color information (spectral color curve) is generated for each color target and loaded into the ink formulation software, which then formulates an ink recipe, based on the ink database.

The ink recipe is sent to the ink dispenser, which prepares the ink batch for production. After quality control the ink goes to press. During the printing stage the Color Measurement System shows color differences, solid density and dot gain values via operator touch screen.

This novelty enables sharing of the color data in a common format, processing it between in-line color control and off-line color measurement technology. By that this solution offers potential for significantly reduced start-up waste and time savings.

High speed digital UV inkjet press

The EFI^{TM} VUTEk$^{®}$ HS100 Pro introduced a high-speed digital UV inkjet press that represents an alternative to analog equipment. Color accuracy and consistency are delivered by innovative imaging technology driven by Orion OS software platform which also ensures demanded productivity and digital efficiency.

The new technology maximizes ink yield for a lower cost of ownership with true grayscale heads and ink delivery system. Innovative Pin & Cure technology delivers precise ink lay down ensuring better image quality and high production speeds, gloss control, increased color gamut, wider output capability and fewer artifacts. By that the downtime, waste and loss potential with sophisticated warning and detection systems are minimized. The new UV inkjet printer can run multiple jobs at one time with multi-queue, eliminating the need of ripping large layout files.

Printer enables prints of maximum 3.2 m with the speed of more than 100 boards per hour or up to 60 boards per hour in Point-Of-Purchase (POP) mode. Prints can be made on different materials as styrene, polypropylene, paperboard and cardboard.

The system is managed with new user interface Digital-Fiery-Frontend application. EFI claims that this kind of technology offers more than the alternative classic offset and screen printing production.

Innovations in RIP software

Version 8 of the RIP software family is now available after successful developing innovative software solutions for the Large Format Printing (LFP), Commercial Printing, Prepress, Packaging Printing and Industrial Printing.

The new versions of the Productionserver, Filmgate, Plategate and Proofgate can be characterized by their high degree of flexibility and scalability. Furthermore, all versions are future-proof ensured by the Value Pack software maintenance program. All RIP software products version 8 are running on the new Adobe PDF Print Engine 3.1. Additionally, the Container+ function, the Linearization Assistant, the Profiler Module PFM and the CutServer has been updated with new functions. The selection of the media size is now open to user defined changes.

All RIP software base products can be configured freely via the Output Management Sets OMS and the optional Ink Saver Module. With Ink Saver, the total amount of ink can be reduced by up to 30 percent without a visible loss of quality, which also means a reduction of the total printing costs.

The complete new version of the software was developed by ColorGate.

Near future is in 4D printing

With 3D printing already highly developed, another innovation is currently under research. It is 4D printing, using adaptive composite materials. The initial concept, introduced by the Massachusetts Institute of Technology, enables production of objects fixed in one shape that can later be changed to take a new shape. Further research, conducted at the University of Colorado, proved that shape memory polymer fibres incorporated into the composite materials obtained in traditional 3D printing, can result in the production of shape-changing objects.

The ability of shape memory polymer fibres to generate desired shape changes of the composite material is based on different physical mechanisms and on how the architecture of the fibres is designed, including their location, orientation and other factors.

A printer for greater productivity

As a result of cooperation with many partners and results from every team involved, Riso developed the new ComColor line of printers. The focus was shifted from the high speed to higher productivity, reducing total downtime and increasing output print speed by developing new features and functions designed to reduce hassle and to complete tasks sooner.

In order to make the most of the high speed and allow achieving the desired results across a range of purposes, a new technical approach is adopted, thus expanding the lineup of peripherals, matching inks and papers, and developing printer's image processing capabilities. These new functions will allow customers to raise their work productivity to even greater levels.

Innovative content and color proofing solution

Created primarily for newspaper production, UB2 certified softproof, hardproof and annotation workflow for documents and images is possible to use within a web browser, throughout the design, review and production process.

PrintPreview UB2 introduces a new document review capability to WebShare UB2, with hierarchical annotation support. This works with multiple-page documents, as well as with all major image formats. It tracks document versions and allows signing off all ready-to-go documents.

Annotation author, project title, file status and due date information is indexed and searchable via the Spotlight search feature. The annotation access e-mail includes a URL which allows users to jump directly from the e-mail into the annotation board by entering the credentials only. UB2 was developed and released by German Helios and certified by FOGRA.

Inks of new generation

The Universal Mild Solvent - UMS inks are Mutoh's fourth generation of mild solvent inks. Developed specially for Mutoh's roll-to-roll sign & display wide-format printer portfolio and intended for long-term outdoor and indoor applications.

These new UMS inks offer an outdoor UV resistance of up to three years without lamination and are virtually odor-free, therefore no ventilation is necessary in the workplace.

According to the manufacturer, these new inks offer a unique and innovative blend of high quality durable color pigments, a new mix of resins and, a new combination of milder and non-aggressive solvents. They offer an unrivalled color gamut and a unique media gloss preservation. The inks are suited for both high quality and high speed printing, hardly requiring any printer maintenance and are fixed and dried at regular heating temperatures.

UMS inks consist of highly durable color pigments grinded to nano-size level (< 100 nm), a new blend of milder and non-aggressive solvents and a new mix of resins. The resin mix enables optimum adhesion of the pigments to the substrate, creating a uniform and smooth ink layer.

Available in bulk 1 liter bottles and 440 ml cassettes UMS inks are suited for a wide variety of applications like posters, banners, backlit signage, fleet signage, wall coverings, fine art, POS displays, billboards, building announcements, etc.

Flame resistant and magnetic papers

Antalis has presented two new special papers - Maine M1, a flame retardant coated paper, and MagneCote, a paper with magnetic properties.

MAINE M1 FLAME RETARDANT

Flame retardant paper Flamstop Maine M1 is a media primarily intended for commercial applications. The basis for the development of the paper has been ensuring public safety. The product is, with an increased whiteness, gloss and a reduced drying time, suitable for offset, screen and digital printing of large formats.

The Maine M1 range is available with FSC® certificate for one side (120 g) and both sides (185, 250, 320 and 770 g) printing.

DIGITAL MAGNECOTE

MagneCote presents a true magnetic paper with printing properties of ordinary paper and acting's of a magnet. It was designed by combining premium coated papers with a proprietary magnetic layer (20 % paper and 80 % magnetic coating). It can be used by all printing technologies - offset, inkjet, HP and dry toner.

MagneCote paper is suitable for any usual post-press operation: it can be saddle stitched, perfect bound, foil stamped, embosssed, die cut, scored, folded or perforated.

Bookshelf

New New Media

Media ecologist and communicator scholar, Paul Levinson presents in the second edition of his book New New Media, a keen insight into the effects of computer-based communication forms. The book documents author's encounters with various contemporary forms including blogging, wikis, podcasts, and social networks like Facebook and MySpace.

Along with a multitude of examples from actual web experience the "new new" media is compared with traditional media. Beside mentioned suggestions on how to widespread adoption of these new forms will affect existing social institutions and attitudes is described.

Besides useful understanding of the mechanics of Twitter, YouTube, Facebook, Wikipedia and other types of new media the book also discusses the impact of this so called "new new" media - Foursquare, Pinteres, WikiLeaks, Anonymous, and Goggle+ on our society. Levinson's down-to-earth discussion of the "new new" media is an effective introduction to the impact of cyberspace structures and institutions on our current media environment.

New New Media
Author: Paul Levinson
Publisher: Pearson; 2nd edition (2012)
ISBN: 978-0205865574
240 pages
160 x 13 x 234 mm
Paperback

3D Printing for Artists, Designers and Makers: Technology Crossing Art in Industry

This work presents the research by Stephen Hoskins and his 3D team at the Centre for Fine Print Research, world leaders in the development of techniques for 3D printing in ceramics. The book explains how the creative industries are directly interfacing with this new technology and how it is changing the practices of many artists and designers across the globe. A selection of case studies of leading practitioners in their respective disciplines reveals this fascinating process in action. It includes a history of 3D printing, from its origins in aerospace to its current, diverse applications in bio-medics and Formula One racing, through to furniture design and jewelry.

It is a fascinating investigation into how the applied arts continue to adapt to new technologies and is suitable for academics and 3D print users from both the arts and science backgrounds, as well as artists, designers, those in creative industries and anyone who has an interest in new technological developments.

3D Printing for Artists, Designers and Makers:
Technology Crossing Art in Industry
Authors: Steve Hoskins, Stephen Hoskins
Publisher: Bloomsbury Visual Arts (2014)
ISBN: 978-1408173794
144 pages
Paperback

Digital Textile Design

Authors: Melanie Bowles, Ceri Isaac

Publisher:
Laurence King Publishers,
2nd edition (2012)
ISBN: 978-1780670027
192 pages
277x18x213 mm
Paperback

Digital Textile Design is covers all the fields of digital designing and printing.

The book examines how designers can access this technique, looking at the work of those currently exploring its possibilities, and provides an insight into the technology involved in digital textile printing.

The work is an excellent guide for students and practitioners who are starting the exploration of digital printing.

Development of high quality printing paper using kenaf fibers

Kenaf (Hibiscus cannabinus), an alternative fiber for the pulp and paper industries in developing countries

Author: Alireza Ashori

DEVELOPMENT OF HIGH QUALITY PRINTING PAPER USING KENAF FIBERS
LAMBERT

Publisher:
Lap Lambert Academic
Publishing (2010)
ISBN: 978-383832112X
172 pages
150x10x226 mm
Paperback

Kenaf, a fast-growing plant, grows wild, but is also widely cultivated. Its fibre is applied for many industrial purposes.

In this research the investigation of suitability of Malaysian cultivated kenaf fibers in the production of high quality printing paper was investigated. Beside the chemical, morphological and pulping properties of different fractions of kenaf, the production of bleached pulp using environmentally-friendly method, TCF, is presented. Beside mentioned, the conventional ECF bleaching sequences are also used to compare the results with the TCF sequences.

The research also presents polymer deposition, surface topography and printability. The overall conclusion is that whole stem kenaf is an attractive raw material that is suitable for use in the production of high quality printing paper in areas where forest resources are limited or inadequate.

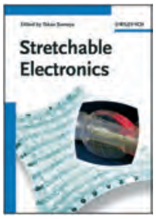
Stretchable Electronics

Requirements for technology are daily becoming more creative and innovative and the field of electronics is leading the way to more advanced appliances. In this book the undisputable knowledge of some of the most outstanding scientists in the field is assembled explaining how to make electronics stretchable.

Therefore the book focuses on gathering and evaluating the materials, designs, models and technologies that enable the fabrication of fully elastic electronic devices which can sustain high strain. It provides a comprehensive review of the specific applications that directly benefit from highly compliant electronic, including transistors, photonic devices and different sensors.

In addition to stretchable devices, the topic of ultraflexible electronic is treated, highlighting its upcoming significance for the industrial-scale production of electronic goods for the consumer market.

The book is divided into four parts which cover theory, materials and processes, circuit boards, devices and applications.



Stretchable Electronics
Editor: Takao Someya
Publisher: Wiley-VCH (2013)
ISBN: 978-3527329786
484 pages
178x25x249 mm
Hardcover

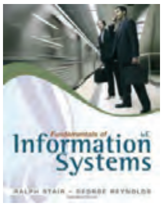
Fundamentals of Information System

This 6th edition presents the latest concise overview of fundamentals of information system. In nine short chapters core principles of information systems are explored, providing readers an engaging overview of the todays discipline and the rapidly changing role of today's information system professional. More than 50 new references in each chapter bring the latest information system topics and examples to the forefront, while new opening vignettes, cases and special interest features clearly demonstrate the emphasis today's organizations place on innovation and speed.

The book provides new discussions on multimedia in business, application developments for new devices, cloud computing, forecasting, and other issues reshaping information systems today. This edition presents some of the most recent research on virtual communities and work structures, including the knowledge on how social networking sites are assisting virtual teams and how companies are effectively using virtual organizational structures with mobile workers.

This book enables the readers the complete overview that they need to function more effectively as workers, managers, decision makers, and organizational leaders in business today.

Fundamentals of Information System
Authors: Ralpf Stair, George Reynolds
Publisher: Cengage Learning; 6th edition (2011)
ISBN: 978-0840062185
504 pages
188x20x246 mm
Paperback


Fundamentals of Information Systems
RALPH STAIR · GEORGE REYNOLDS

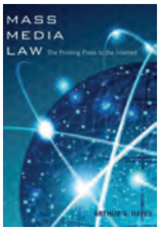
Mass Media Law: The Printing Press to the Internet

With the development of digital media, the law and regulations are now the dynamic legal territory. Mass Media Law presents a textbook designed to introduce students to the panoply of legal theories raised by the Internet revolution as well as those supporting traditional media. The book takes a historical approach beginning with the printing press and the telegraph and proceeding to the digital technologies of today, such as social media and search engines.

Concepts such as defamation, broadcast regulation, privacy, and free expression are covered along with new media legal theories including Internet exceptionalism, cyber libertarianism, and digital speech and democratic culture. These are mostly introduced to explain why traditional theories such as First Amendment medium-specific analysis, common carriage and network neutrality are just as relevant today as they were in the early twentieth century.

In order to help readers develop critical reasoning skills, each chapter opens with a highly readable real-world vignette and goes on to identify and explain legal doctrines and tests. Key passages from court opinions are highlighted, and each chapter closes with a list of online media law resources and thought-provoking questions, including legal hypotheticals, to give readers a solid understanding of the area in question.

Mass Media Law: The Printing Press to the Internet
Author: Arthur S. Hayes
Publisher: Peter Lang International
Academic Publishers (2013)
ISBN: 978-1433107566
304 pages
251 x 18 x 175 mm
Paperback



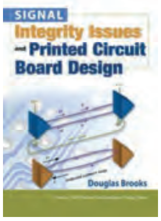
Signal Integrity Issues and Printed Circuit Board Design

With the rapidly developing electronics, there is a need for a basic, comprehensive text that explains how to successfully design boards for any high-speed application problems and their solutions.

Following an easy-to-understand electronics primer for every PCB designer, author offers practical, real-world solutions for every important signal-integrity problem.

The book offers an insight into essential electronic concepts, EMI principles and controls, controlling signal reflections, power systems stability and conditioning and many more topics.

This book covers even more design rules, specific recommendations, examples, illustrations, and diagrams.



Signal Integrity Issues and
Printed Circuit Board Design
Author: Douglas Brooks
Publisher: Prentice Hall (2012)
ISBN: 978-0133359473
432 pages
180 x 23 x 229 mm
Paperback

Environmentally Benign Approaches for Pulp Bleaching

Author: Pratima Bajpai

Publisher:
Elsevier Science;
2nd edition (2012)
ISBN: 978-0444594211
416 pages
191 x 30 x 234 mm
Hardcover

Globally increasing of pulp and paper production will also continue to increase in the near future. To cope with the increasing demand, an increase in production and improved environmental performance is needed as the industry is constantly pressured to reduce environmental emissions to air and water.

New pulping and bleaching technology, strict effluent regulations, environmental pressure groups and new market demands have had a considerable influence on modern bleaching practices.

This book offers updated information on environmentally benign approaches for pulp bleaching, which can help solve the problems associated with conventional bleaching technologies.

The Digital Print:

Preparing Images in Lightroom and Photoshop for Printing

Author: Jeff Schewe

Publisher: Peachpit Press;
1st edition (2013)
ISBN: 978-0321908452
336 pages
Paperback

Renowned photographer, educator, and author Jeff Schewe presents targeted chapters on digital printing from Lightroom and Photoshop and shares his expert techniques for optimal output and fine-art reproduction.

A companion to *The Digital Negative: Raw Image Processing in Lightroom, Camera Raw, and Photoshop*, this book teaches how to take already perfected images and optimize them for the best quality final printing.

Readers can learn about different printer types and principles of color management so that expected results can be achieved. Strategies on proofing, sharpening, re-solution, black-and-white conversion, and workflow, as well as on identifying the attributes that define a perfect print are also presented.

Bookshelf

Academic dissertations

Doctoral thesis - Summary

Author:
Silva König

Speciality field:
recycled paper, print quality, durability

Supervisor:
Diana Gregor Svetec

Co-supervisor:
Tadeja Muck

Defended:
September 2013, at University of Ljubljana, Faculty of Natural Sciences and Engineering, Ljubljana, Slovenia

Contacts:
silva.könig@ntf.uni-lj.si

Properties and print quality of recycled papers

The aim of the thesis was to investigate graphic papers made of recycled and virgin fibres, their print quality and durability. In the research study, two 100 % recycled papers and two wood-free papers made of virgin fibres were included. The basic, mechanical, surface and optical properties of papers, their fibre components and surface structure were determined. The papers were examined with the UV/VIS and FT-IR spectroscopy. To determine print quality, the investigated graphic papers were printed on with the electrophotographic printing technology, using dry toners, and later evaluated with image analysis. Beside mentioned the research of accelerated aged papers and prints was made with treatments using increased temperature and humidity and irradiation of xenon arc lamp. After treatments some paper properties, surface structure, and the UV/VIS and FT-IR spectroscopy were determined once again. The print durability of offset and electrophotography which uses liquid toners was evaluated with spectrophotometric measurements and with new chrominance histogram method.

The results showed that some mechanical properties ensure better resistance of recycled papers due to their larger fibre fibrillation and more filled structure between fibres. Due to small ink particles that were left in the pulp and smaller content of optical brighteners, recycled papers had lower brightness. The prints on recycled papers established slightly smaller colour gamut. Nevertheless, the results showed that recycled papers are suitable for everyday office use. During accelerated ageing, the tensile strength of investigated papers reduced. The results proved slightly better light fastness of optical properties and degree of polymerization for recycled papers. Moreover, the results showed hydrolytic and oxidative degradation of cellulose, this being most noticeable at the uncoated recycled paper. The results also revealed different colour stability of prints where the effect depended on the pigment type in ink and paper coating. The type of fibres did not influence print durability. A new method of assessing the light fastness of prints with the image chrominance histogram quantification proved very useful and led to valuable results.

Doctoral thesis - Summary

Author:
Robin Abderrahmen

Supervisor:
Naceur Belgacem

Co-supervisor:
Didier Chaussy

Defended:
December 2012 at PAGORA Grenoble, France

Contacts:
Naceur.Belgacem@pagora.grenoble-inp.fr
Didier.Chaussy@pagora.grenoble-inp.fr

Design of self-adhesive labels by microencapsulation adhesive

The main objective of this study was to prepare innovative silicone liner-free labels. This can be achieved by the adhesive "self protection" thanks to its incorporation into microcapsules. This allows the preparation of "dry labels" gluing under the application of a pressure, which induces the rupture of the microcapsules, thus releasing the core material, a pressure sensitive adhesive.

The first step was to analyse three water-based PSA in view of their encapsulation. Then, the most suitable adhesive was microencapsulated by coacervation (using biopolymer as shell) and by in situ polymerisation. Two other encapsulation processes (spray-cooling and spray-drying) were also carried out and were compared with the two former processes. Coating colour formulations were prepared with spray-cooling microcapsules (the most adhesive ones). Coating trials were carried out by blade coating and by screen printing.

Compatibility between microcapsules and the label making process, using a flexographic printing press, was determined. Finally, the mains characteristics of the prepared innovative products (adhesion, application pressure) were compared to industrial self-adhesive homologues, and found that they could be suitable for the preparation of silicon liner-free envelops and stamps.

The Journal of Print and Media Technology Research will publish summaries of high quality academic thesis within the scope of the journal. Short summaries should be sub-mitted to <journal@iarigai.org> by the thesis supervisor. Information on type and field of the thesis, author, supervisor, date and university of defense or presentation, as well as on how the full thesis can be obtained must be provided.

Events

Media Port

Berlin, Germany
7 to 9 October 2013

One of the events related to the IFRA Expo & conference and the International Newsroom Summit is the Media Port Agenda, that will offer strategies and solutions for efficient processes, multi-channel publishing, community building and paid content for the newsroom.

The Media Port agenda includes the following topics:

- *Community building:* Video, blogger networks and social media at Digital First Media in the USA, Hürriyet in Turkey, Warum Verlag in Germany and others;
- *Monetising digital:* Paywall solutions at Sanoma in Finland, Regionalmedien Austria, Dépêche Du Midi in France and others;
- *Efficient workflows in the newsroom:* CMS and more at Axel Springer in Germany, Shaw Media in the USA, Direktpress in Sweden and others;
- *Responsive design:* Best practices at Irish Times, Polaris Media in Australia, DuMont Net in Germany and others;
- *Hyperlocal publishing:* The local approach of Umuntu Media in South Africa, Johnston Press in the UK, WAZ NewMedia in Germany and others.

Print Fair

Mumbai, India
6 to 10 November 2013

In collaboration of Printweek India and Goethe Institut, this five-day event to showcase the depth and breadth of top print work and print ideas in India. The intention of the Print Fair is to represent the full spectrum of the best of Indian print work.

PRINT FAIR
AT
MAX MUELLER BHAVAN MUMBAI
6-10 November 2013
Presented by PrintWeek

At Print Fair, the best print samples from the PrintWeek India Awards with a range of diverse interests, will be laid out and documented for the benefit of the visitors, which will include students, designers, print production professionals, print buyers, brand owners, etc. Besides this, Print Fair will feature an array of print workshops, panel discussions, and print-related film screenings on each day of the show. A B2B workshop, which is also slated at the fair, will have print companies discuss print innovations and trends for the audience from design community.

Contec

Frankfurt, Germany
8 October 2013

As a prelude to the Frankfurt International Book Fair (9 to 13 October 2013) the Frankfurt Academy will introduce Contec, a conference designed to encourage sinergy among all members of the publishing community. This new conference format will certainly contribute to the rapidly changing face of the Book Fair. Contec will reflects an industry in which publishing and technology are already inextricably intertwined. What was once seen as a "clash of cultures" is well on its way to achieving a symbiotic balance, and Contec will be exploring this "new publishing experience". Participants will be introduced to new voices from startups and small innovative companies, self-published authors, trade and academic publishers, library specialists from all over the world. With no formal presentations, the sessions will be focused on dialogue and exchange on a range of essential topics, including data, the future of bookselling, the implications of self-publishing on the industry, responsive design, metadata, rights, distribution, trends, forecasts, partnerships with tech sector and startups. This engaging program is intended to offer the tools needed to move the business forward.

Paperex

New Delhi, India
24 to 27 October 2013

Paperex is a renowned business platform in form of international exhibition and conference for the pulp, paper and allied industries. A presence of more than 500 leading exhibitors from over 40 countries is expected, with more than 30 000 visitors from around the globe. A high level technical conference will be organized in addition, to serve the industry as a forum for experts to share their knowledge and experience.

Pamex

New Delhi, India
14 to 17 November 2013

This year's edition of Pamex will feature new developments in machinery and materials and will showcase generation next technologies from worldwide solution providers across categories from pre-press to post-press and everything in between. Over 30 000 visitors from 25 counties are expected to visit Pamex and check more than 300 exhibitors.

WAN-IFRA is continuing a successful series of specialized regional events.

Digital Media Latinoamérica 2013

Bogota, Colombia
30 and 31 October 2013

The Latin American news industry is taking its own digital path. Building on the success of our Digital Media events in Europe, Asia and India, which attracted over 800 senior publishing executives Digital Media Latinoamérica (DML) is a chance to network with like-minded colleagues from across the publishing industry who strive for that essential added edge in today's hyper-competitive publishing environment. It is designed for the most forward thinking top, mid-level and executive media managers.

Digital media Asia

Kuala Lumpur, Malaysia
12 to 14 November 2013

Fifth Digital Media Asia conference will gather approximately 300 participants from 32 countries, representing over 150 media companies and suppliers to the news publishing industry. Some of the hot topics to be discussed will include: Content issue; Implementing Big Data; Integrating print and online metrics; The power of publisher's advertising partnerships; Start-ups meet publishers; Monetizing online video; From newspaper to broadcaster; New ways to increase mobile revenue; Latest trends on digital and e-commerce.

Print & Packtech World Expo-2013

Bangalore, India
27 to 30 September 2013

Printing and Packaging has experienced revolutionary changes. It's amazing and breathtaking to compare anything and everything that's happening today. With what used to happen may be 10 years back, it is no longer the world of unpacked things, virtually everything on earth can be packed. Packaging is no longer meant for the mere protection of a product but has become critical market differentiator.

Print & Pack World Expo-2013 will be the premier exhibition for the printing and packaging industries, raising awareness and augment Indian printing and packaging capabilities, spurring both sectors to new heights.

Print & Pack World Expo-2013 will be tailor-made to reflect the market's evolutionary path driven by new needs, competition, technology, channels and other developments. The exhibition will have a special focus on the integration of processes in printing and packaging and its supply chain as this will eventually lead both industries to claim market leadership.

The exhibition will present excellent opportunities for the industries to evaluate its investing decisions in printing and packaging.

Paper-Me 2013

Cairo, Egypt
28 to 30 November

Paper Middle East Exhibition is the premier event for Middle East & North Africa's pulp, paper, tissue, paper board making and products industries. It is the only international paper exhibition sponsored and supported by all paper industry-related governmental departments and nationwide trade associations of Egypt.

The most important specialized fair for the global paper, paperboard, tissue and converting industry, where all nations and customers from all over the world will be gathered, as well as the Middle East and Arabian countries with high purchasing and decision making power in one place.

The Middle East is a fast growing market with a population of over 450 million people. It is reported recently that the Middle East is the future of paper production and consumption and import in the world. The Middle East and North Africa offer a wealth of opportunities for the paper, board, tissue manufacturers and suppliers.

Paper Middle East Exhibition (PaperME) continues to grow going from strength to strength. It posted a significant increase of 30 % in the number of exhibitors and 19.67 % in the number of visitors in comparison to the last edition of the event. PaperME is considered to be the natural venue of machinery manufacturers interested in international expansion for their products throughout the MENA region.

Call for papers

The Journal of Print and Media Technology Research is a peer-reviewed periodical, published quarterly by iarigai, the International Association of Research Organizations for the Information, Media and Graphic Arts Industries.

Authors are invited to prepare and submit complete, previously unpublished and original works, which are not under review in any other journals and/or conferences.

The journal will consider for publication papers on fundamental and applied aspects of at least, but not limited to, the following topics:

- Printing technology and related processes

  Conventional and special printing; Packaging, Fuel cells and other printed functionality; Printing on biomaterials; Textile and fabric printing; Printed decorations; Materials science; Process control

- Premedia technology and processes

  Color reproduction and color management; Image and reproduction quality; Image carriers (physical and virtual); Workflow and management

- Emerging media and future trends

  Media industry developments; Developing media communications value systems; Online and mobile media development; Cross-media publishing

- Social impacts

  Environmental issues and sustainability; Consumer perception and media use; Social trends and their impact on media

Submissions for the journal are accepted at any time. If meeting the general criteria and ethic standards of scientific publishing, they will be rapidly forwarded to peer-review by experts of high scientific competence, carefully evaluated, selected and edited. Once accepted and edited, the papers will be printed and published as soon as possible.

There is no entry or publishing fee for authors. Authors of accepted contributions will be asked to sign a copyright transfer agreement.

Authors are asked to strictly follow the guidelines for preparation of a paper (see the abbreviated version on inside back cover of the journal). Complete guidelines can be downloaded from:

http://www.iarigai.org/publications/

Papers not complying with the guidelines will be returned to authors for revision.

Submissions and queries should be directed to:

journal@iarigai.org or office@iarigai.org

Guidelines for authors

Authors are encouraged to submit complete, original and previously unpublished scientific or technical research works, which are not under review in any other journals and/or conferences. Significantly expanded and updated versions of conference presentations may also be considered for publication. In addition, the journal will publish reviews as well as opinions and reflections in a special section.

Submissions for the journal are accepted at any time. Papers will be considered for publishing if meeting the general criteria and ethic standards of the scientific publication. When preparing a manuscript for JPMRT, please strictly comply with the journal guidelines, as well as with the ethic aspects. The Editorial Board retains the right to reject without comment or explanation manuscripts that are not prepared in accordance with these guidelines and/or if the appropriate level required for scientific publishing cannot be attained.

A - General

The text should be cohesive, logically organized, and thus easy to follow by someone with common knowledge in the field . Do not include information that is not relevant to your research question(s) stated in the introduction.

Only contributions submitted in English will be considered for publication. If English is not your native language, please arrange for the text to be reviewed by a technical editor with skills in English and scientific communication. Maintain a consistent style with regard to spelling (either UK or US English, but never both), punctuation, nomenclature, symbols etc. Make sure that you are using proper English scientific terms.

Do not copy substantial parts of your previous publications and do not submit the same manuscript to more than one journal at a time. Clearly distinguish your original results and ideas from those of other authors and from your earlier publications - provide citations whenever relevant. For more details on ethics in scientific publication, please consult:

http:// www.elsevier.com/ethicguidelines.

If it is necessary to use an illustration, diagram, table, etc. from an earlier publication, it is the author's responsibility to ensure that permission to reproduce such an illustration, diagram etc. is obtained from the copyright holder. If a figure is copied, adapted or redrawn, the original source must be acknowledged.

Submitting the contribution to JPMTR, the author(s) confirm that it has not been published previously, that it is not under consideration for publication elsewhere and - once accepted and published - it will not be published under the same title and in the same form, in English or in any other language. The published paper may, however, be republished as part of an academic thesis to be defended by the author. The publisher retains the right to publish the printed paper online in the electronic form and to distribute and market the Journal (including the respective paper) without any limitations.

B - Structure of the manuscript

Title: Should be concise and unambiguous, and must reflect the contents of the article. Information given in the title does not need to be repeated in the abstract (as they are always published jointly).

List of authors: i.e. all persons who contributed substantially to study planning, experimental work, data collection or interpretation of results and wrote or critically revised the manuscript and approved its final version. Enter full names (first and last), followed by the present address, as well as the e-mail addresses.

Separately enter complete details of the corresponding author - full mailing address, telephone and fax numbers, and e-mail. Editors will communicate only with the corresponding author.

The title of the paper and the list of authors should be entered on a separate cover page (numbered as 0). Neither the title nor the names of authors can be mentioned on the first or any other following page.

Abstract: Should not exceed 500 words. Briefly explain why you conducted the research (background), what question(s) you answer (objectives), how you performed the research (methods), what you found (results: major data attained, relationships), and your interpretation and main consequences of your findings (discussion, conclusions). The abstract must reflect the content of the article, including all the keywords, as for most readers it will be the major source of information about your research. Make sure that all the information given in the abstract also appears in the main body of the article.

Keywords: Include three to seven relevant scientific terms that are not mentioned in the title. Keep the keywords specific. Avoid more general and/or descriptive terms, unless your research has strong interdisciplinary significance.

Abstract and keywords should be entered on a separate page, numbered as page 1. Do not continue with the main body of the text, regardless of the possible empty space left on this page.

Introduction and background: Explain why it was necessary to carry out the research and the specific research question(s) you will answer. Start from more general issues and gradually focus on your research question(s). Describe relevant earlier research in the area and how your work is related to this.

Methods: Describe in detail how the research was carried out (e. g. study area, data collection, criteria, origin of analyzed material, sample size, number of measurements, equipment, data analysis, statistical methods and software used). All factors that could have affected the results need to be considered. Make sure that you comply with the ethical standards, with respect to the environmental protection, other authors and their published works, etc.

Results: Present the new results of your research (previously published data should not be included). All tables and figures must be mentioned in the main body of the article, in the order in which they appear. Do not fabricate or distort any data, and do not exclude any important data; similarly, do not manipulate images to make a false impression on readers.

Discussion: Answer your research questions (stated at the end of the introduction) and compare your new results with the published data, as objectively as possible. Discuss their limitations and highlight your main findings. At the end of Discussion or in a separate section, emphasize your major conclusions, specifically pointing out scientific contribution and the practical significance of your study.

Conclusions: The main conclusions emerging from the study should be briefly presented or listed, with the reference to the aims of the research and/or questions mentioned in the Introduction and elaborated in the Discussion.

Introduction, Methods, Results, Discussion and Conclusions - as the scientific content of the paper - represent the main body of the text. Start numbering of these sections with page 2 and continue without interruption until the end of Conclusions. Number the sections titles consecutively as 1, 2, 3 ..., while subsections should be hierarchically numbered as 2.1, 2.3, 3.4 etc. Use Arabic numerals only.

Note*: Some papers might require different structure of the scientific content. In such cases, however, it is necessary to clearly name and mark the appropriate sections.*

Acknowledgments: Place any acknowledgments at the end of your manuscript, after conclusions and before the list of literature references.

References: The list of sources referred to in the text should be collected in alphabetical order on a separate page at the end of the paper. Make sure that you have provided sources for all important information extracted from other publications. References should be given only to documents which any reader can reasonably be expected to be able to find in the open literature or on the web. The number of cited works should not be excessive - do not give many similar examples. Responsibility for the accuracy of bibliographic citations lies entirely with the authors.

Please use only the Harvard Referencing System. For more information consult, e. g., the referencing guide at:

http:// libweb.anglia.ac.uk/referencing/harvard.htm.

List of symbols and/or abbreviations: If non-common symbols or abbreviations are used in the text, you can add a list with explanations. In the running text, each abbreviation should be explained the first time it occurs.

Appendix: If an additional material is required for better understanding of the text, it can be presented in the form of one or more appendices. They should be identified as A, B, ... etc., instead of Arabic numerals.

Above sections are supplementary, though integral parts of the Scientific content of the paper. Each of them should be entered on a separate page. Continue page numbering after Conclusions.

C - Technical requirements for text processing

For technical requirement related to your submission, i.e. page layout, formatting of the text, as well of graphic objects (images, charts, tables etc.) please see detailed instructions at http:// www.iarigai.org/publications/journal.

D - Submission of the paper and further procedure

Before sending your paper, check once again that it corresponds to the requirements explicated above, with special regard to the ethic issues, structure of the paper as well as formatting. Once completed, send your paper as an attachment to: **journal@iarigai.org**. You will be acknowledged on the receipt within 48 hours, along with the code under which your submission will be processed. The editors will check the manuscript and inform you whether it has to be updated regarding the structure and formatting. The corrected manuscript is expected within 15 days. At the same time the first (or the corresponding) author will be asked to sign and send the Copyright Transfer Agreement.

Your paper will be forwarded for anonymous evaluation by two experts of international reputation in your specific field. Their comments and remarks will be in due time disclosed to the author(s), with the request for changes, explanations or corrections (if any) as demanded by the referees. After the updated version is approved by the reviewers, the Editorial Board will consider the paper for publishing. However, the Board retains the right to ask for a third independent opinion, or to definitely reject the contribution. Printing and publishing of papers once accepted by the Editorial Board will be carried out at the earliest possible convenience.

3-2013

Journal of Print and Media Technology Research

A peer-reviewed quarterly

The journal is publishing contributions in the following fields of research:

- Printing technology and related processes
- Premedia technology and processes
- Emerging media and future trends
- Social impacts

For details see the Mission statement inside

JPMTR is listed in

Index Copernicus International

PiraBase and PaperBase
(by Smithers Pira)

Submissions and inquiries
journal@iarigai.org

Subscriptions
office@iarigai.org

More information at
www.iarigai.org/publications/journal

Publisher
The International Association of Research
Organizations for the Information, Media
and Graphic Arts Industries
Washingtonplatz 1
D-64287 Darmstadt
Germany

Printed in Croatia by Narodne Novine, Zagreb