Pp. 282–287 in *Proceedings of the 2003 Symposium on Applications and the Internet Workshops (SAINT 2003 Workshops)*. IEEE Computer Society, Los Alamitos, California, 2003.

Bayesian Analysis of Online Newspaper Log Data

Hannes Wettig, Jussi Lahtinen, Tuomas Lepola, Petri Myllymäki and Henry Tirri
Complex Systems Computation Group (CoSCo), Helsinki Institute for Information Technology (HIIT)
University of Helsinki & Helsinki University of Technology
P.O.Box 9800, FIN-02015 HUT, Finland
{firstname}.{lastname}@hiit.fi

Abstract

In this paper we address the problem of analyzing web log data collected at a typical online newspaper site. We propose a two-way clustering technique based on probability theory. On one hand the suggested method clusters the readers of the online newspaper into user groups of similar browsing behaviour, where the clusters are determined solely based on the click streams collected. On the other hand, the articles of the newspaper are clustered based on the reading behaviour of the users. The two-way clustering produces statistical user and page profiles that can be analyzed by domain experts for content personalization. In addition, the produced model can also be used for on-line prediction so that given the user cluster of a person entering the site, and the page cluster of an article of a newspaper, one can infer whether or not the user will have a look at the page in question.

1. Introduction

Over the last decade a vast amount of web log data has been collected all over the world, mainly by companies hoping to find out more about the browsing behaviour of the visitors of their site. For many reasons rigorous statistical analysis of this data has proven to be a much more complicated problem than originally expected. Therefore web log data analysis still in many cases means only simple counting of page hits. However, site design based on statistical analysis of actual user behavior data — let alone development of personalized content — is a goal that cannot be achieved without much more elaborate statistical modeling techniques.

In this paper we address this problem area and focus on the analysis of web log data collected at a typical online newspaper site. One of the problems typical to newspaper data is that the contents of the site are frequently changed, normally on a daily basis. In addition, the user population is also dynamically changing in time: new users register in every day, and some of the old users stop using the service.

For solving the above problems, we propose a *two-way clustering technique* based on probability theory. On one hand the suggested method clusters the readers of the online newspaper into user groups of similar browsing behaviour, where the clusters are determined only based on the click streams collected. On the other hand, the articles of the newspaper are clustered based on the reading behaviour of the users. The goal in this two-way clustering is to find general statistical user and page profiles so that given the user cluster of a person entering the site, and the page cluster of an article of a newspaper, we will be able to predict probabilistically whether or not the user will have a look at the page in question. By moving from models of individual users and pages to models of user and page groups we are able to cope statistically with individual users and pages not seen before, i.e., new registered users and new content pages.

The data used in this research is described in more detail in the next Section. In Section 3 we present a general discussion on the modeling problem at hand, and contrast our approach to previous related work. The probabilistic two-way clustering scheme is described in Section 4: the criterion for choosing between different clusterings is defined in Section 4.1, and the search algorithm for finding good clusterings is given in Section 4.2. The observations made during the empirical validation of the suggested approach are summarized in Section 5.

2. The data

For this research we were provided user web log data from the *Iltalehti Online* site (http://www.iltalehti.fi), a major Finnish online newspaper. The data contained site visits within a survey period of one month. During data cleaning the data had been pre-pruned so that only users with at least three non-trivial sessions during the survey period were part of the sample. In our context, we refer to a session as non-trivial, if it includes at least one article page and not merely headline browsing on the front page. After each day, the entire content of the site changes, however, the basic structure (sections) of the site is preserved. None of the changes occur during the day.

The raw data given to us contains the usual click streams, matched to an (anonymous) user ID (for legal reasons, names and other identifiers were removed from the data). Furthermore, for each ID we were given the demographic profile information (consisting of data about education, work, age and sex) users were asked to give of themselves when registering for the service. Finally, for each URL, the following meta data was provided: the headline the article was referred by, the section it was linked to and stored in, the category it had been assigned to by a journalist (using a standard news categorization), whether the article contained foreign or domestic news and whether or not there was a direct link from the front page to the article. This information was partly missing or unreliable (uncategorized pages, users misinforming facts about themselves etc.).

We pre-processed the click streams to obtain an $n \times m$ hit matrix $D = (d_{ji})$ with a line for each user U_j ($j = 1..n$) and a column for each page P_i ($i = 1..m$) during the survey period. The entries in the matrix were either $+$ (the user viewed the page), $-$ (did not view) or $?$ (did not visit on the day in question). In the case of this study, the data D was stored as a 2045×643 matrix. Thus there were $n = 2045$ users who during the month in question had at least three non-trivial sessions, and altogether there had been $m = 643$ individual articles on the site. Viewing time was not taken into account in this study, since as also pointed out by Shahabi et al. [11], it was found the be too unreliable.

3. Modelling the data

The objective of this research was to construct methods for extracting statistical information that could be applied in site design and/or in developing personalized content. Site design requires a deeper understanding of the users in general, while personalization requires understanding the behaviour of a single user. The goal was that the developed methods could be used, for example, for pointing out the (subjectively) most interesting headlines to a user on the front page. Another way of providing personalized content is a "top ten list" of articles, directly listing the pages the user is most likely to visit. The personalized content can be dynamically updated on-line according to the clicks made by the user as the session proceeds. In order to provide this type of information, we need a model that generalizes from one user to another and reveals general trends. In other words, we need to simplify the data as much as possibly, while losing as little information as possible.

For solving the modeling problem, Cadez et al. propose a first-order Markov model (see [1]) for user session clustering. This did not seem favourable here for the following resons. First, the order of the clicks (links visited) depend very strongly on the physical structure of the site, and does not necessarily reflect the mental state of the user. Secondly, as already discussed, the click stream data is most unreliable with respect to the paths taken through the site. This is mainly due to missing cache and back-button hits, as demonstrated in e.g. [11].

The information we want to model and predict is whether or not a user will view a certain page, given he or she has entered the site. We call this variable "View"; it depends on both the page and the user. The full Bayesian Network representation of this situation is shown in Figure 1 (for a general introduction to Bayesian Network models, see e.g. [9, 4]).

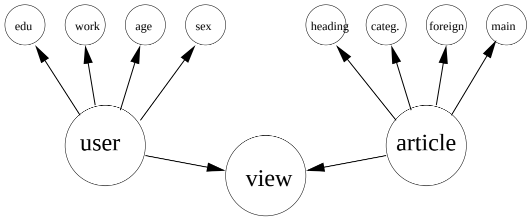


Figure 1. Full Bayes Net representation of the domain, including the meta information.

With this approach, we have to learn multiple independent causes (user *and* page) of the observed data (view); this is sometimes referred to as *Factorial Learning*, see e.g. [3]. Commonly in this situation, the data is assumed to be normally distributed, and in particular to be continuous in the first place. Under these assumptions, Gharamani [3] used an EM type algorithm, but noticed that the exact computation was intractable.

In this work we are in a different, potentially more complicated situation, since our data is binary, and hence we want to model it using a binomial distribution. From this perspective, Figure 1 is just another (Bayesian) representation of the data (hit matrix), and clearly this "model" is much too complicated, as it contains one parameter for each matrix entry (even where it is missing, i.e. "?"), so it does not generalize at all.

In our first modeling approach, we used a *Finite Mixture* model (see e.g. [6]). For this we took a different perspective, and first identified each page with its (hand-picked) category, and then created one variable for each category. All sessions of one user could now be treated as one vector, with entries for each category. These entries in turn were multiple values of the form "user U viewed r pages of category C out of s offered to him", referring to r positive ($+$) and $s - r$ negative ($-$) events. We clustered these vectors by adding a single hidden variable which was learned by an EM algorithm (see e.g. [7]).

However, there were 119 categories appearing during the one-month time span, which seemed to be too many for our purposes. As a consequence, we obtained very many rather small clusters that did not provide satisfactory insight into typical browsing behaviour of the users — this model generalized very poorly.

In our second approach we replaced each page by its section, of which there are five in our case. Although Huang et al. [5] strongly advocate this idea and even declare that "*sessions are essentially sequences of categories*", it is clear that this naturally means losing a lot of information. However, this approach has often worked well in practice (see e.g. [1]).

The effect of this approach was opposite to the preceding: five attributes seemed to be few and we were able to extract only some basic information from the data. As a result, we obtained only few clusters with little variance in terms of the demographic user profiles, providing only vague hints about individual user behaviour.

For solving the above problems, Oyanagi et al. propose a matrix clustering in a situation very similar to ours [8]. However, the suggested "ping-pong algorithm" has some serious drawbacks. First of all, the method cannot handle the missing data present in our hit matrix. Second, in each iteration i, the method produces a pair of clusters $(C^i_{row}, C^i_{column})$, such that the rows in C^i_{row} are similar *only* wrt. the columns in C^i_{column} and vice versa, while what we actually want is clusters whose members are similar to each other as a whole. Furthermore we do not want to fix the number of row clusters to equal the size of column clusters, but learn both from the data. Finally, the ping-pong algorithm does not cluster the complete matrix, but there remains a "left-over" or "garbage" cluster, of which there can be said nothing at all.

In the following we pick up the idea of simultaneous row and column clustering, but put it in a more sound theoretical framework based on (Bayesian) probability theory.

4. Two-way clustering

We suggest to model the domain by a Bayesian Network containing two hidden (i.e. unobserved) variables, one grouping the users into clusters of similar browsing behaviour and one grouping the pages into clusters of similar visitor sets. As we shall see, this approach provides us with a number of page types that lies inbetween the two numbers mentioned above, and which is adaptive wrt. to the data size.

Figure 1 implies that the relation between users and pages is essentially symmetric. Consequently, as well as clustering the users based on the pages they view, we can cluster the pages based on their support, i.e., based on the group of users they were viewed by. This type of Bayesian clustering in both horizontal and vertical directions (of the matrix) has been suggested earlier in the context of text analysis [10].

In this paper we suggest Bayesian clustering of rows and columns *simultaneously*, which – to our knowledge – has not been done before. From Figure 1 we arrive at the Bayesian Network depicted in Figure 2 by adding two hidden variables, clustering users and pages, respectively. Note that the direction of the arcs between X and "user" and between Y and "article" is arbitrary for reason of network equivalence as discussed in [9].

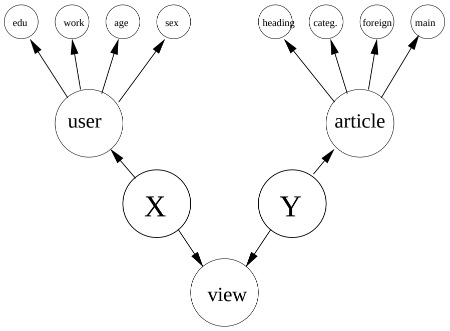


Figure 2. Bayes Net with two hidden clustering variables X and Y added.

The full model contained $n \cdot m$ parameters, while the hidden variables reduce the size of the probability table at the node "view" to $K_x \cdot K_y$, where $K_x := |X|$ and $K_y := |Y|$ denote the number of the model's user and page clusters respectively. Naturally, this number will be a lot smaller, i.e. $K_x \cdot K_y << n \cdot m$. On the other hand, the model can express the data well only if users (pages) within each cluster are similar wrt. the "view"-variable, given the page (user) clustering. In the following we denote the "view"-variable by C (for *class*).

Two central questions are now the following: How many clusters are there in the data? How to cluster the data? These issues will be addressed in the following two subsections.

4.1 The model selection criterion

Let M denote the Bayesian model shown in Figure 2. For model selection criterion, we suggest to optimize the *complete data evidence* (cf. [6]) $P(C, X, Y|M)$, which factorizes as

$$P(C, X, Y|M) = P(X|M)P(Y|M)P(C|X, Y, M), \quad (1)$$

where the first two terms limit clustering complexity and the latter term is the actual data modelling part. Maximizing (1) thus means looking for the most probable data completion in the unobserved variables X and Y. Note that, in slight abuse of notation, X is a vector of length n, Y has length m. Furthermore we disregard missing entries ("?") in C, which we view as a binary variable, and thus C has length equal to the combined number of entries "+" and "-" in the hit matrix. We denote these numbers by h^+ and h^- respectively, their sum simply by $h := h^+ + h^-$.

Assuming the (complete) data to be *i.i.d.* (independent and identically distributed), the terms in (1) can be computed efficiently as follows:

$$P(X|M) = \frac{\prod_k \Gamma(h_k^x + \mu_k^x)}{\Gamma(n + \sum_k \mu_k^x)}, P(Y|M) = \frac{\prod_l \Gamma(h_l^y + \mu_l^y)}{\Gamma(m + \sum_l \mu_l^y)},$$

$$P(C|X,Y,M) = \prod_{k,l} \frac{\Gamma(f_{kl}^+ + \sigma_{kl}^+)\Gamma(f_{kl}^- + \sigma_{kl}^-)}{\Gamma(f_{kl} + \sigma_{kl}^+ + \sigma_{kl}^-)}, \quad (2)$$

where the *sufficient statistics* are as follows: h_k^x is the number of users in (user) cluster k, h_l^y the number of pages in (page) cluster l, and f_{kl}^+ and f_{kl}^- the numbers of positive resp. negative page view events for users in cluster k and pages in cluster l, their sum being $f_{kl} := f_{kl}^+ + f_{kl}^-$. The μ's and σ's are *priors* or *hyperparameters*, which we simply set to one; this choice is sometimes referred to as the *uniform prior*. Note that Eq. (2) also gives the *prequential* data likelihood [2], which corresponds to the situation where the data is available to us one session/page click at a time and we have to predict in real-time, while updating the sufficient statistics.

In the above, M implicitly included the cluster sizes K_x and K_y. But since we do not know how many clusters there should be, we cannot fix these numbers in advance. Separating the cluster sizes from the model family M, we arrive at

$$P(C,X,Y,K_x,K_y|M) = P(K_x|M)P(X|K_x,M) \cdot P(K_y|M)P(Y|K_y,M)P(C|K_x,X,K_y,Y,M). \quad (3)$$

The terms conditioned on the cluster sizes are still given by (2) (where K_x and K_y were implicit), but now we have two new terms that correspond to priors on the cluster sizes. In order to be proper, both priors should sum to one over all possible cluster sizes (here: the integers). In this study we simply set $P(K_x|M) := 2^{-K_x}$ and $P(K_y|M) := 2^{-K_y}$. In our experiments we also tried different priors, but that did not seem too affect the final results.

We can now formulate our goal: we wish to find user and page clusterings X and Y of a priori unknown sizes K_x and K_y, so that the following objective function is maximized:

$$P(C,X,Y,K_x,K_y|M) = 2^{-K_x-K_y} \frac{\prod_{k=1}^{K_x} \Gamma(h_k^x + 1)}{\Gamma(n + K_x)} \cdot \frac{\prod_{l=1}^{K_y} \Gamma(h_l^y + 1)}{\Gamma(m + K_y)} \prod_{k,l} \frac{\Gamma(f_{kl}^+ + 1)\Gamma(f_{kl}^- + 1)}{\Gamma(f_{kl} + 2)}, \quad (4)$$

where the sufficient statistics depend on X and Y and we view the priors to be part of our model choice.

4.2 The search algorithm

Optimizing (4) is clearly a hard task. In this study we used a stochastic greedy algorithm for finding good solutions; optimality cannot be expected. The algorithm performs, in each iteration, the following $n + m + 2(K_x + K_y)$ operations in random order:

- *for each user* (resp. *page*) assign him (it) to a user (page) cluster such that (4) is maximized
- *for each user (page) cluster* assign *all* vectors therein to other clusters such that (4) is maximized and delete the emptied cluster; discard if (4) decreases
- split *each user (page) cluster* in two as described below; discard if (4) decreases

The algorithm terminates when no changes were made during one complete iteration. The search starts from an initial random clustering. Interestingly, the size of the initial clustering did not seem to correlate with the final clustering sizes.

During the splitting operation we only look at the row (resp. column) vectors in the cluster to be splitted and assign them to two new clusters independently of all the other row (column) vectors. First, we create an ordering of the vectors in question. We do this by creating two empty clusters A and B, which will gradually be filled by alternatingly assigning to them one vector at a time. The first vector to be assigned to cluster A is chosen completely at random. From there on, we look at A and B as a two-component mixture model, such that for each vector d not yet assigned to one of the clusters we get a probability distribution over A and B by normalizing $(P(A|d), P(B|d)) := c(P(d|A), P(d|B))$, where $c = (P(d|A) + P(d|B))^{-1}$. In each step we now assign to B (resp. A) that vector d which maximizes $P(B|d)$ ($P(A|d)$), i.e. which gives maximum probability to the cluster in consideration. Ties are broken at random.

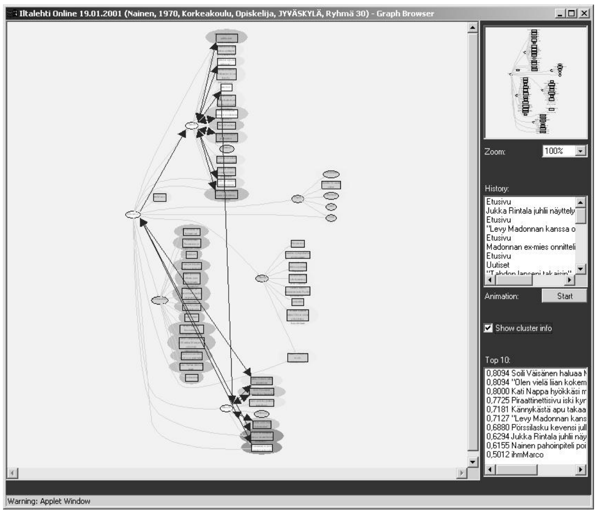
Next we merge the two new clusters A and B in an ordering where we first have all the vectors in A in the order they were assigned, followed by the vectors of cluster B in reversed order of assignment. Then we find a splitting point in the ordering maximizing the objective function (4). The resulting two clusters are taken as the final outcome of the split operation, and we delete the old original cluster.

Although this algorithm does not find the globally optimal user and page clusterings (because of randomness in the splitting operation it might not even find a local optimum), we find that it provides quite usable results. It seems that these operations hinder premature convergence to a large degree: the operations suggested enable the algorithm to perform 'quantum jumps' to new interesting areas of the search space. Also note that the cluster deletion and splitting operations enable the algorithm to choose the clustering sizes K_x and K_y fully automatically.

5. Practical observations

The algorithm grouped the 2045 users into about 60 clusters, with little variance between the single runs of our stochastic algorithm. The clusters itself were found to make a lot of sense by the domain experts, affirming known patterns of behaviour and revealing new and useful information about reader browsing. Also, the user clusters followed the demographic user profiles quite nicely. As this meta data was not used in the clustering, we see this outcome as evidence for both the appropriateness of our model and the truthfulness of most registered users of the Iltalehti Online site.

User clusters can be viewed as prototypes of the form "a small group of male students, who like football and who are not interested in celebrity gossip". This information can be used to provide personalized information to the user. We have also developed a tool that visualizes user sessions along with the predictions of our model and a top ten list of suggestions to the user (Fig. 3 shows an animated path through the site, history and top ten predicted articles.).


: Iltalehti Online 19.01.2001 (Nainen, 1970, Korkeakoulu, Opiskelija, JYVÄSKYLÄ, Ryhmä 30) - Graph Browser
Zoom:
100%
History:
Animation:
Start
☑ Show cluster info
Top 10:
Warning: Applet Window

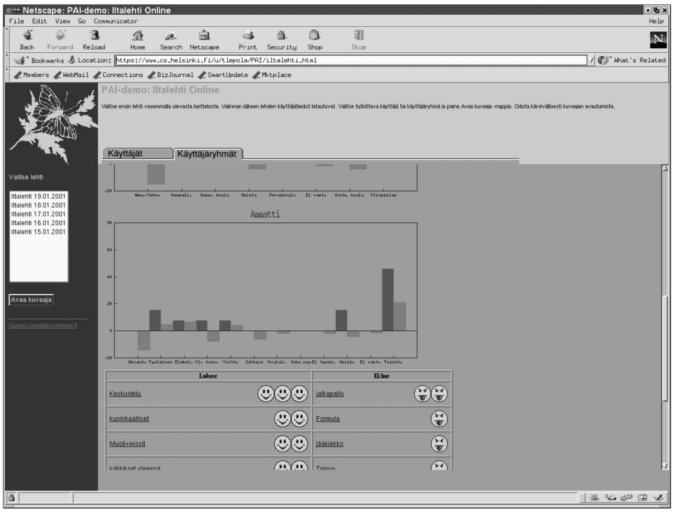
History:

Etusivu
Jukka Rintala juhlii näyttely
Etusivu
"Levy Madonnan kanssa o
Etusivu
Madonnan ex-mies onnitteli
Etusivu
Uutiset
"Tahdon lapseni takaisin"

Top 10:

0,8094 Soili Väisänen haluaa M
0,8094 "Olen vielä liian kokem
0,8000 Kati Nappa hyökkäsi m
0,7725 Piraattinettisivu iski kyr
0,7181 Kännykästä apu takaa
0,7127 "Levy Madonnan kans
0,6880 Pörssilasku kevensi jull
0,6294 Jukka Rintala juhlii näy
0,6155 Nainen pahoinpiteli poi
0,5012 ihmMarco

Figure 3. Visualization of a user session.



Location: https://www.cs.helsinki.fi/u/tlepola/PAI/iltalehti.html

PAI-demo: Iltalehti Online

Valitse ensin lehti vasemmalla olevasta luettelosta. Valinnan jälkeen lehden käyttäjätiedot latautuvat. Valitse tutkittava käyttäjä tai käyttäjäryhmä ja paina Avaa kuvaaja -nappia. Odota kärsivällisesti kuvaajan avautumista.

Valitse lehti:

Iltalehti 19.01.2001
Iltalehti 18.01.2001
Iltalehti 17.01.2001
Iltalehti 16.01.2001
Iltalehti 15.01.2001

Lukee		Ei lue	
Keskustelu	😊😊😊	jalkapallo	☹☹
kuninkaalliset	😊😊	Formula	☹
Muoti+missit	😊😊	jääkiekko	☹
julkkikset yleensä	😊😊	Talous	☹

Figure 4. Visualization of a user cluster.

In addition to this, our tool visualizes the user clusters (see Figure 4). Profiles are given as bar charts, common likes and dislikes of the group are pictured by a range of one to three happy or sad smilies.

Whenever a user has visited the site before, we have information on the browsing behaviour enabling us to assign him or her to a suitable cluster. For new users we can guess a cluster based on the given profile, or infer it dynamically based on the current page clicks as the session proceeds. Similarly, we can assign the pages of the newspaper of a new day to clusters initially based on available meta data and make more educated choices after we have observed some site traffic.

The 643 pages were grouped into about 30 clusters, again with little variance. The page clustering followed closely the sections of the site, but split these further down into 'subsections', just as we had hoped. The few exceptions found here were quite intuitive: e.g., a "sports" article talking about an ice-hockey player's marriage was grouped together with pages in section "celebrities".

The results suggest that categorizing the pages manually is not obligatory any more, as the suggested approach provides us with sensible categories automatically. The number of produced clusters is adaptive, growing with increasing data availability. Moreover, the page clusters span over all editions of the newspaper of the whole month, so that the user clustering is not biased by the days on which a user visited the site.

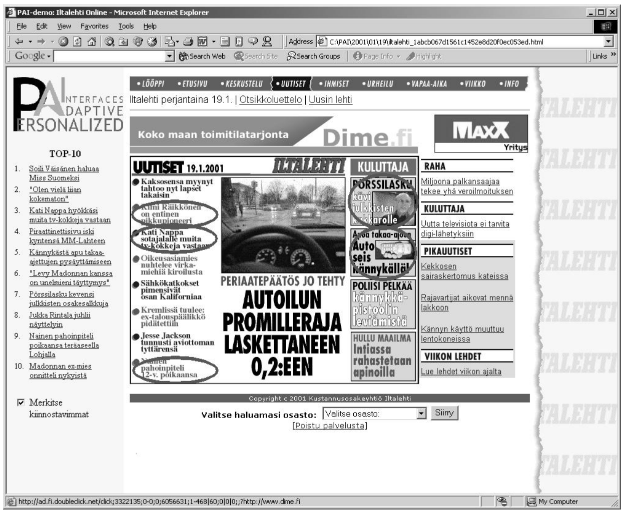
6. Conclusions

In this work our goal was to build a statistical model that can be used for two distinctive purposes: for analyzing off-line the user clusters of a typical online newspaper site, and for predicting on-line the future behavior of the users entering the site. For achieving this goal, a two-way probabilistic clustering approach was suggested, based on the idea of simultaneously clustering the individual users and the daily pages into probabilistic groups or profiles. The potential clusterings were compared by using a theoretically justified Bayesian modeling criterion (marginal likelihood) which in this case could be computed efficiently. For finding good clusterings, a simple stochastic greedy search algorithm turned out to be a feasible solution for practical purposes.

The best statistical model constructed from the data, i.e., best with respect to the criterion used, contained approximately 60 user clusters and 30 page clusters. The off-line analysis of the clusters, performed by domain experts, showed that the clusters found made intuitive sense so that the model revealed new and interesting regularities in the data. This can be said both for the user clusters and the page clusters. The former distinguish themselves in terms of browsing behaviour by construction, but in addition also in terms of the profiles given by the users themselves. The page clusters in turn contain articles of similar content regardless of the newspaper issue they appeared in.

The on-line predictions provided by the statistical model were validated empirically by using special-purpose software developed in the project. An example of a personalized WWW interface made possible by the approach suggested is given in Figure 5: for each user, the system first computes for each page the probability with which the user will have a look at the page during the current session. The 10 most probable pages of today's paper are then shown in the "top ten list" on the left. In addition to this, the headlines of the predicted most relevant articles are highlighted by a red circle.

We leave it for future research to actually implement a personalized user interface of this type. It remains an open question how to evaluate the success of such a system, since it will inevitably change the user's browsing behaviour and along with it the nature of the log data acquired.


PAI-demo: Iltalehti Online - Microsoft Internet Explorer

Address C:\PAI\2001\01\19\iltalehti_1abcb067d1561c1452e8d20f0ec053ed.html

File Edit View Favorites Tools Help
Google
Search Web
Search Site
Search Groups
Page Info
Highlight
Links
PAI
INTERFACES
ADAPTIVE
PERSONALIZED
• LÖÖPPI • ETUSIVU • KESKUSTELU • UUTISET • IHMISET • URHEILU • VAPAA-AIKA • VIIKKO • INFO
Koko maan toimitilatarjonta
Dime.fi
MAXX
Yritys
ILTALEHTI
Siirry
My Computer

Iltalehti perjantaina 19.1. | Otsikkoluettelo | Uusin lehti

TOP-10

1. Soili Väisänen haluaa Miss Suomeksi
2. "Olen vielä liian kokematon"
3. Kati Nappa hyökkäsi muita tv-kokkeja vastaan
4. Piraattinettisivu iski kyntensä MM-Lahteen
5. Kännykästä apu takaa-ajettujen pysäyttämiseen
6. "Levy Madonnan kanssa on unelmieni täyttymys"
7. Pörssilasku kevensi julkkisten osakesalkkuja
8. Jukka Rintala juhlii näyttelyin
9. Nainen pahoinpiteli poikaansa teräaseella Lohjalla
10. Madonnan ex-mies onnitteli nykyistä

☑ Merkitse kiinnostavimmat

UUTISET 19.1.2001

- Kaksosensa myynyt tahtoo nyt lapset takaisin
- Kimi Räikkönen on entinen pikkupioneeri
- Kati Nappa sotajalalle muita tv-kokkeja vastaan
- Oikeusasiamies nuhtelee virkamiehiä kiroilusta
- Sähkökatkokset pimensivät osan Kaliforniaa
- Kremlissä tuulee: ex-talouspäällikkö pidätettiin
- Jesse Jackson tunnusti aviottoman tyttärensä
- Nainen pahoinpiteli 12-v. poikaansa

PERIAATEPÄÄTÖS JO TEHTY
AUTOILUN PROMILLERAJA LASKETTANEEN 0,2:EEN

KULUTTAJA
PÖRSSILASKU kävi julkkisten kukkarolle
Apua takaa-ajoon Auto seis kännykällä!
POLIISI PELKÄÄ kännykkäpistoolin leviämistä
HULLU MAAILMA Intiassa rahastetaan apinoilla

RAHA
Miljoona palkansaajaa tekee yhä veroilmoituksen

KULUTTAJA
Uutta televisiota ei tarvita digi-lähetyksiin

PIKAUUTISET
Kekkosen sairaskertomus kateissa
Rajavartijat aikovat mennä lakkoon
Kännyn käyttö muuttuu lentokoneissa

VIIKON LEHDET
Lue lehdet viikon ajalta

Copyright c 2001 Kustannusosakeyhtiö Iltalehti

Valitse haluamasi osasto: Valitse osasto:

[Poistu palvelusta]

http://ad.fi.doubleclick.net/click;3322135;0-0;0;6056631;1-468|60;0|0|0;;?http://www.dime.fi

Figure 5. An example of a personalized interface made possible by the probabilistic modeling approach suggested.

Acknowledgements. This research has been supported by the National Technology Agency and the Academy of Finland.

References

[1] I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Visualization of navigation patterns on a web site using model based clustering. In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD),* Boston, MA, pages 280–284. ACM Press, 2000.

[2] A. Dawid. Prequential analysis, stochastic complexity and Bayesian inference. In J. Bernardo, J. Berger, A. Dawid, and A. Smith, editors, *Bayesian Statistics 4*, pages 109–125. Oxford University Press, 1992.

[3] Z. Ghahramani. Factorial learning and the em algorithm. In *Advances in Neural Information Processing Systems 7*, pages 617–624. MIT Press, 1995.

[4] D. Heckerman. Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1(1):79–119, 1997.

[5] Z. Huang, J. Ng, D. W. Cheung, M. K. Ng, and W. Ching. A cube model for web access sessions and cluster analysis. In *WEBKDD 2001,* San Francisco, CA, 2001.

[6] P. Kontkanen, P. Myllymäki, and H. Tirri. Comparing Bayesian model class selection criteria by discrete finite mixtures. In D. Dowe, K. Korb, and J. Oliver, editors, *Information, Statistics and Induction in Science*, pages 364–374, Proceedings of the ISIS'96 Conference, Melbourne, Australia, August 1996. World Scientific, Singapore.

[7] P. Kontkanen, P. Myllymäki, and H. Tirri. Constructing Bayesian finite mixture models by the EM algorithm. Technical Report NC-TR-97-003, ESPRIT Working Group on Neural and Computational Learning (NeuroCOLT), 1996.

[8] S. Oyanagi, K. Kubota, and A. Nakase. Application of matrix clustering to web log analysis and access prediction. In *WEBKDD 2001,* San Francisco, CA, 2001.

[9] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, CA, 1988.

[10] F. C. N. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. In *Meeting of the Association for Computational Linguistics*, pages 183–190, 1993.

[11] C. Shahabi, F. Banaei-Kashani, and J. Faruque. A reliable, efficient, and scalable system for web usage data acquisition. In *WEBKDD 2001,* San Francisco, CA, 2001.