The Development and Application of an Evaluation Methodology for Person Search Engines

Roland Brenneke
Information Science
University of Hildesheim
Marienburger Platz 22
Germany
roland.brenneke@gmx.de

Thomas Mandl
Information Science
University of Hildesheim
Marienburger Platz 22
Germany
mandl@uni-hildesheim.de

Christa Womser-Hacker
Information Science
University of Hildesheim
Marienburger Platz 22
Germany
womser@uni-hildesheim.de

ABSTRACT

This paper presents a user oriented evaluation methodology for comparing person search services on the Web. Many established system oriented methods from information retrieval cannot be applied to this domain. Our user oriented methodology is applied to a test comparing the person search engines yasni, pipl.com and 123people. The user study with over 30 participants led to relevant results. The coverage of data object types within the person search engine results is quite different. Especially the amount of pictures and social media network entries which are presented by the systems and which are perceived by the test users differ greatly. The results also revealed a tendency to judge people more positively when more information was found.

1. INTRODUCTION

Person search engines are important specialized search services on the Web. These systems consult other services for information about a person and integrate it in one interface. They can be regarded as meta search services or one point stops for personal information. Mostly, they are tailored for normal people and not for celebrities and other famous people. As such, it is different from named entity search in general.

Especially in the Web 2.0 and its ease of publishing content on the Web, many people deposit much information about them or content they created in various sites. Users need to have the proper information competence to foresee the consequences of such behavior. Often, users are advised not to publish too much information. Online reputation management becomes an important issue. On the side of the users, social networks and person search services lead to information ethical considerations about the use of personal information.

Searching on information about others is a very frequent information need and a reason for using a search service. According to Google Trends, the most popular person search services receive over 200,000 hits per day. However, 90% of the users do not rely on person search engines but they use general Web search or go directly to social networks to find out about people. Nevertheless, 10% is still a significant share and hit rates for person search engines are constantly high. In addition, many of these searches may have a high impact. Many recruiters use person search engines for checking on candidates.

A questionnaire study among 548 enterprises was published in 2010 [5]. This Social Media HR Report 2010, revealed that in 2009 over 59% of the companies have used the internet to check on applicants. Almost 10% had already turned down an application because of information on the Web. Companies who do not use the Web for checking on applicants` state that lack of time and ethical questions are the main reasons not to do so [5].

An international study showed that this behaviour is more widespread in the US than in European countries [3]. Interviews with decision makers in German companies revealed that they are well aware of the potential of retrieving applicant information [11].

The use of person search engines for job applicants is only one potential usage scenario; however, it is a very prominent one. Other than that, there are many reasons for why a user would want to search for a person. And despite the use of a named entity in the search, the information need is rather vague and can be rephrased with “Find out something about person X”.

The success of a person search engine depends on many factors. Person search engines are meta services which extract results from a large variety of different online media. The presentation of these results in the user interface is an essential factor for the success of the search service. If a result is far down on the result page and the user never scrolls there, potentially relevant items cannot be found. That means that the search capability is only one success factor for person search engines. Consequently, our experiment was designed as a user test. We intended to evaluate the user experience and the success with the tool person search engine and neither specific system components nor absolute retrieval performance.

2. RELATED WORK

The evaluation of retrieval systems is central in information retrieval research because the system performance cannot be predicted. The most influential retrieval evaluation methodology is called the *Cranfield* paradigm. Information retrieval research has adopted an evaluation scheme which tries to ignore subjective differences between users in order to be able to compare systems and algorithms. The user is replaced by a prototypical and constant user. Relevance judgments are provided by domain experts [8, 10].

EuroHCIR 2011. The 1st European Workshop on Human-Computer Interaction and Information Retrieval. July 4th 2011. Newcastle, UK

Cranfield evaluations have often been criticised for several reasons. The main objections come from advocates of user oriented studies. The search situation of users depends on many individual and contextual factors which can only be captured in user experiments [6]. The real user experience and the success in a real world situation cannot be measured with the laboratory style experiments based on the *Cranfield* paradigm [12].

Person search engines have a higher chance to succeed than general purpose search services. The retrieval with named entities is known to be easier than searches without names entities [9]. The selection of a person search engine hints the type of result. Consequently, synonymy between names and words are a smaller problem than in general purpose search engines. Synonymy between names, on the other hand, is a big challenge for person search engines.

3. METHODOLOGY

The balance between control and realism is a challenge for each experiment. For the presented study, we chose a user experiment to test person search engines because an approach purely dedicated to retrieval power does not mirror the user experience for person search engines well. It is necessary to limit the realism in a user experiment in order to allow comparison across participants in the test. We selected a job applicant scenario in order to make the experiment interesting for the users. Applicant search is a very prominent usage type. The method was successful in making the experiment attractive. The test users liked the experiment very much and through word of mouth, more applicants wanted to register for the experiment than were needed.

The selection of persons for the task defines the content for the test. It seemed necessary to identify people for whom much information can be found on the Web. If there were no videos, working results like presentations or social network entries, then the performance of the person search engine could not be tested with our experiment. So even if the persons selected are not representative in terms of amount of online information for the whole population or all persons who are indexed in a person search service it increases the validity of the test to select persons with a large amount of online information.

Three people were carefully selected who had similar qualifications. For them, a job profile was developed which was given to the participants together with the names of the people. The users were asked to search for these people who would be interviewed for the position and check if they were appropriate. The job description and the name of each applicant were given to the test persons. Each of the candidates was well qualified for the job but had one negative aspect in his online data. One was an advocate of nuclear power and the job was for offered by an alternative energy company. The second applicant was a serial entrepreneur who portrayed himself on Facebook in pictures with attractive women and sports cars. The third applicant had party photos online where he could be seen smoking cigarettes and he considered himself as lazy in one social network while he had a very business oriented self image in another social network.

Obviously, such a scenario has some limitations. Person search engines need to disambiguate between people with the same name. We decided to choose people who are not ambiguous in order to have the same difficulty for each person. Such issues are evaluated in the system oriented campaign WEPS [1].

We selected people who had posted a large amount of information about themselves in the network. Again, this was done to obtain similar and comparable difficulty for the three test cases. Three person search engines were selected for the comparative test. We chose yasni, pipl.com and 123people because they were very popular at the time of the study according to Google trends. All three companies claim that they exploit only information available on the public Web.

4. STUDY

Students of the University of Hildesheim were recruited through a mailing list of students. Participation was voluntarily and no gratification was given. None of the participants had a computer science background. They all were frequent Internet users and had searched for people before but only 10% had used a person search engine before. The others use Google or social networks to find information on people.

The issue of relevance is always a crucial one in information retrieval evaluation. In our study, any item could contribute to the full picture of the applicant. Despite the clearly defined scenario, it remains vague which information is needed and what type of information is useful. It is difficult to assign relevance to items or even weights to categories. The user interfaces of the person search engines present the items in categories like e.g. social network entries or videos.

A questionnaire study [7] showed that users search mainly for the following items in the order presented when retrieving information about a specific person:

- Contact information
- Profile on a social network
- Photo
- Information about professional accomplishments or interests

The most frequently researched item, contact information does not apply for our scenario because the persons had sent a letter of application. The next two most frequent items are included. The fourth item is rather vague as some of the other items following as far as the categories of person search engines are concerned. As a consequence, the data available does not justify the assignment of weights to some items. In our study, all clicks on items were scored equally. The results will also show which of the items were most popular. The time per applicant was limited to 10 minutes. The entire experiment took 45 minutes on average including the pre- and post questionnaire.

One search service modified the interface after the first two tests. So it was necessary to eliminate three test sessions from the results and recruit further test users. This shows that not only the dynamics of the personal data presents a challenge for the test but also the ongoing modifications of the search engine. Overall, 34 took part in the experiment. Due to the problems of a relaunch of one service, we could consider the experiments of 10 users of 123people, 11 users of Pipl and 10 user of Yasni.

Each test person worked with one search engines on all three applicants. This between groups approach was applied was mainly applied to avoid a long learning phase for each of the person search engines. All tests were recorded with appropriate software.

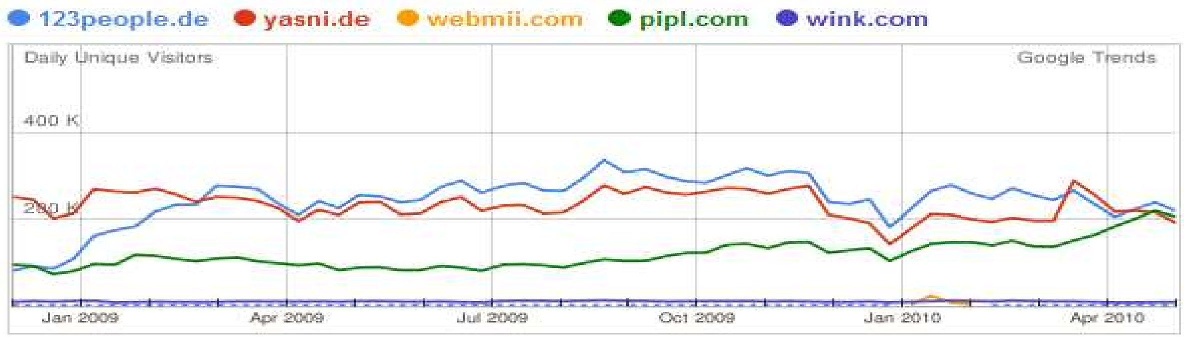

Figure 1: Popularity of person search engines according to Google Trends

5. RESULTS

The result description focuses on the information perceived by users and the performance of the test users in the application task.

The information items clicked by the users were categorized. It can be seen that the services lead to a similar number of clicks when summed up over all users. Each of the services resulted in between 110 to 120 clicks for the ten test persons. In the case of Pipl, 11 test persons were considered. Each engine leads to a sufficient number of entries and has abundant information on the applicants in our scenario. This was a goal of the test design and was accomplished.

The type of information which was encountered was quite different. It can be easily seen, that 123.people facilitates access to photos whereas Pipl leads more users to social network entries. A comparative analysis for the services for the most popular item types is shown in Table 1.

In the post test questionnaire, users were asked about their subjective impression of the service they had used. In the overall satisfaction, 123people was rated highest. For the page structure, pipl received the best grades and the coverage of different business networks yasni was rated as most successful. In the latter case, the finding from the objective click data was confirmed. Further details on the results are provided in [2].

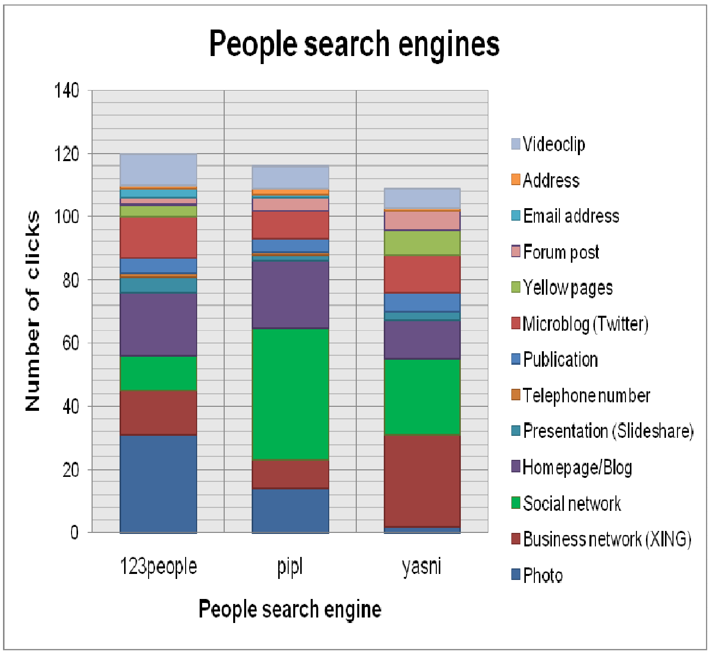

Figure 2: Clicks on items in the three person search engines

Table 1: Comparison of data types encountered

Item	123people	Pipl	Yasni
Photo	+ +	+ –	– –
Business network	–	–	+ +
Social network	–	+ +	+
Homepage/Blog	+	+	+ –
Microblog	+	+ –	+
Yellow pages	+ –	– –	+
Forum post	–	+ –	+
Videoclip	+	+ –	+ –
Publication	Because of a very low number of clicks is no rating possible.		
Presentation			
Email address			
Address			
Phone number			

Perception	
+ +	Excellent
+	Good
+ –	Moderate
–	Poor
– –	Unperceived

For two services, applicant 1 was selected by the majority of the test users. These two services had identified most items for this applicant. For yasni, applicant 2 was chosen as the best applicant despite the fact that the other two services found on average 10 items more for this person. Applicant 3 was given the last place for all three person search services. For each service, he is the applicant with the fewest items. There might be a trend to rate people higher when more information is available online.

6. RESUME

We presented a holistic evaluation methodology for person search engines. The performance of these search services is measured by observing the perception of test users. The test methodology is built on a realistic scenario and use case but it does not cover all the relevant quality aspects of person search engines. The important capability to resolve the ambiguity of names was not dealt with. In future work, it might be promising to develop a performance based test for this task only.

The complete information seeking behaviour and its success is also not measured with our test. In a realistic scenario, people might access the social media networks through a person search engine and continue their search mainly there. This issue could be resolved by observing real behaviour.

In the test, the search engine 123people was the winner. It not only led users to the highest number of items, but it was also subjectively judged to be the best person search engine. However, in several aspects other systems performed better and were judged better. The evaluation showed that the different tools are all based on the freely available data on the Web but that they lead to different results. The most sought items in our test were photos, entries and profiles in social and business networks and personal homepages. Each of the engines exhibited a strength in one of these items, e.g. 123people for photos because they are shown as top results. This is also confirmed by the questionnaire study among American recruiters [7].

For the users who publish information about themselves and who become information providers by doing that the issue of information competence will become more and more important. Personal Online Identity Management is a growing field and several new companies are entering the market.

7. REFERENCES

[1] Artiles, J.; Borthwick, A.; Gonzalo, J.; Sekine, S.; Amigó, E. 2010. WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks. In: *CLEF Working Notes* http://nlp.uned.es/weps/weps-3/papers

[2] Brenneke, R. 2010. *Evaluation von Personensuchmaschinen und Umgang mit persönlichen Daten im Internet*. Master Thesis, University of Hildesheim, Germany. International Information Management.

[3] CrossTab Marketing Services. 2010. *Europäischer Datenschutztag: Studie zur Online Reputation* Trustworthy Computing Group, Microsoft (Hrsg.). http://www.microsoft.com/germany/sicherheit/datenschutzstudie.mspx

[4] Hellmann, R.; Griesbaum, J.; Mandl, T. 2010. Quality in Blogs: How to find the best User Generated Content. In: *13th Intl Conf on Business Information Systems* (BIS 2010) Berlin, 3.-5. May. Berlin et al.: Springer [LNBIP 47] pp. 47-58.

[5] Zur Jacobsmühlen, T. (2010): *Social Media HR Report* 2010 Stepstone.de & HRM.de (eds.). http://www.jacobsmuehlen.de/studie/

[6] Lamm, K.; Greve, W.; Mandl, T.; Womser-Hacker, C. 2010. The Influence of Expectation and System Performance on User Satisfaction with Retrieval Systems. In: Proc *EVIA 2010: The First Intl Workshop on Evaluating Information Access* June 2010 National Institute of Informatics (NII) Tokyo, Japan, June 15-18, http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/EVIA/09-EVIA2010-LammK.pdf

[7] Madden, M.; Smith, A. 2010. *Reputation Management and Social Media: How people monitor their identity and search for others online.* PEW Internet & American Life Project. http://pewinternet.org/Reports/2010/Reputation-Management.aspx

[8] Mandl, T. 2008. Recent Developments in the Evaluation of Information Retrieval Systems: Moving Toward Diversity and Practical Applications. In*: Informatica – An Intl. Journal of Computing and Informatics* vol. 32. pp. 27-38.

[9] Mandl, T.; Womser-Hacker, C. 2005. The Effect of Named Entities on Effectiveness in Cross-Language Information Retrieval Evaluation. In: *Proc 2005 ACM SAC Symposium on Applied Computing* (SAC). Santa Fe, New Mexico, USA. March 13.-17. 2005. pp. 1059-1064.

[10] Robertson, S. 2008. On the history of evaluation in IR. In: *Journal of Information Science* 34(4). pp. 439-456

[11] Schäuble, T.; Griesbaum, J.; Mandl, T. 2009. Mehrwertpotenziale von Online-Social-Business-Netzwerken für die Personalbeschaffung von Fach- und Führungskräften. In: *Informatik 2009 - Beiträge 39. Jahrestagung der Gesellschaft für Informatik e.V.* (GI) Lübeck [LNI P-154] pp. 2166 – 2180.

[12] Tawileh, W.; Mandl, T.; Griesbaum, J. 2010. Evaluation of five web search engines in Arabic language. In: *LWA– Lernen - Wissensentdeckung – Adaptivität*: Proc Workshopwoche GI, Universität Kassel. Workshop Information Retrieval. http://www.kde.cs.uni-kassel.de/conf/lwa10/papers/ir1.pdf