5th SASTech
Iran | Mashhad
May 12 - 17 | 2011
5th Symposium on Advances in Science & Technology

Designing a Meta-Search Engine Considering the User's Field of Interest

Farzaneh Zahmatkesh, zahmatkesh.farzaneh@gmail.com
Hamid Hassanpour, h_hassanpour@yahoo.com
Paper Reference Number: 0602-991
Name of the Presenter: Farzaneh Zahmatkesh

Abstract

Existing meta-search engines return web search results based on the page relevancy to the query, their popularity and content. In addition, they disregard the user's preferences or field of interest. It is necessary to provide a meta-search engine capable of ranking results considering the user's field of interest. Social Networks can be useful to find the users' tendencies, favorites, skills and interests. In this paper we propose MSE, a Meta-Search Engine for document retrieval utilizing social information of the user. In this approach, each user is associated with a user profile that captures his interests available from a social network he or she belongs to. The MSE receives search results from three underlying search engines. It extracts main phrases from the title and short description of each result. Then it clusters the main phrases by self-organizing feature map algorithm. Generated clusters are then ranked on the basis of the user profile. The more similar cluster label to the user's field of interest gets the higher rank. We have compared the proposed MSE against two other meta-search engines. The experimental results show the efficiency and effectiveness of the proposed method.

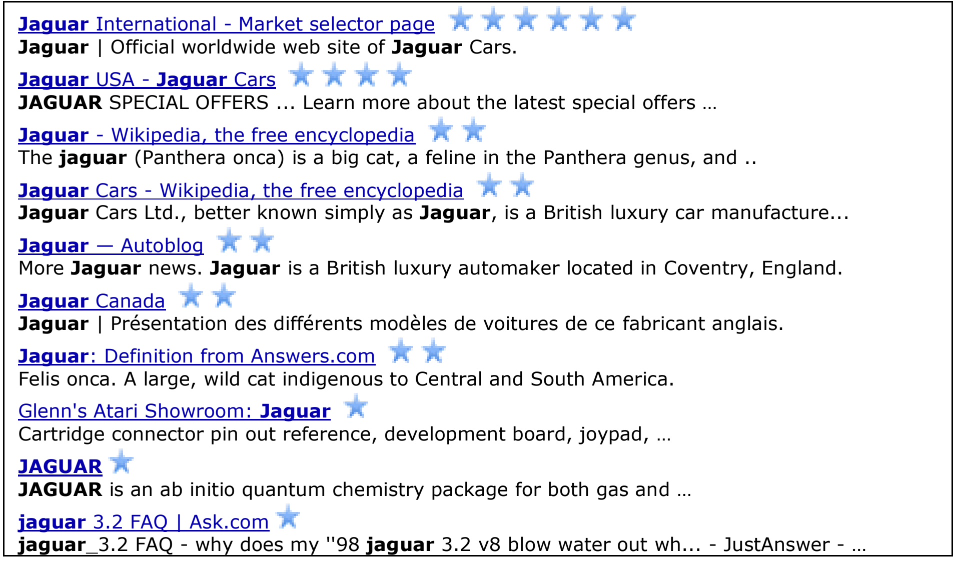
Key words: Clustering, Meta-Search Engine, Ranking, Search Relevance, Social Information

1. Introduction

A meta-search engine is a searching tool that employs the search results of other search engines. Unlike a search engine, a meta-search engine does not maintain its own database of web pages. Instead, search queries are sent simultaneously to several search engines. The gathered results from different resources are then collated to remove duplicates and to rank the results following to the meta-search engine algorithm. Depending on user's search criteria and the meta-search engine itself, results are ranked differently by a variety of methods according to the engine's algorithm. Most meta-search engines rank web pages according to how popular that page is ranked by its sources.

Existing meta-search engines return web search results based on the page relevancy to the query, their popularity and content. In addition, they disregard the user's preferences or field of interest. In the existing meta-search engines, the user usually evaluates the findings one by one to hit upon the proper ones. Different people may look for different resources whilst they utilize the same query or keywords, hence employing a user-independent approach in ranking the web resources may not satisfy the user. Figure 1 shows a search example. There are two users, A and B. User *A is* looking for information on the Mac OS X 10.2, while user B is interested in jaguar cars. They issue the same query, "jaguar" on Ixquick meta-search engine and obtain the default ranked list shown in the Figure 1. This ranking list does not contain needed information for user *A*. What's more, it is not satisfactory for user B, because it contains documents about the jaguar cats too. Thus, a user-aware search system is desirable for improving search effectiveness.

This research explores how information contained in the structure of a social network can enhance search result relevance on a meta-search engine.



Jaguar International - Market selector page ★★★★★★
Jaguar | Official worldwide web site of **Jaguar** Cars.

Jaguar USA - **Jaguar** Cars ★★★★
JAGUAR SPECIAL OFFERS ... Learn more about the latest special offers ...

Jaguar - Wikipedia, the free encyclopedia ★★
The **jaguar** (Panthera onca) is a big cat, a feline in the Panthera genus, and ..

Jaguar Cars - Wikipedia, the free encyclopedia ★★
Jaguar Cars Ltd., better known simply as **Jaguar**, is a British luxury car manufacture...

Jaguar — Autoblog ★★
More **Jaguar** news. **Jaguar** is a British luxury automaker located in Coventry, England.

Jaguar Canada ★★
Jaguar | Présentation des différents modèles de voitures de ce fabricant anglais.

Jaguar: Definition from Answers.com ★★
Felis onca. A large, wild cat indigenous to Central and South America.

Glenn's Atari Showroom: **Jaguar** ★
Cartridge connector pin out reference, development board, joypad, ...

JAGUAR ★
JAGUAR is an ab initio quantum chemistry package for both gas and ...

jaguar 3.2 FAQ | Ask.com ★
jaguar_3.2 FAQ - why does my ''98 **jaguar** 3.2 v8 blow water out wh... - JustAnswer - ...

Figure 1. First ten search results of query "jaguar" on Ixquick.

Traditional approaches to search include scoring documents for relevance based on a set of keywords or using the link structure across documents to infer quality and relevance. These approaches attempt to optimally match keywords to documents with little or no information about the searcher and no information about his or her network. In reality, users are involved in different social communities, and are increasingly engaged in social networks through online services like Facebook, Flickr, and YouTube. The social network is a unique reflection of the user, and can be used to find the users' tendencies, favorites, skills and interests.

In this paper, we propose a meta-search engine named MSE. The MSE uses a social information-based approach of search, where each user corresponds to a profile containing his or her field of interest. The objective is to leverage on social information so that the documents can be ranked based on it. A possible solution to this problem is to online cluster search results into different categories, and to enable users to identify their required category at a glance. In order to evaluate MSE, we conduct a comparative performance study among MSE and two other meta-search engines, and our results show that MSE is an effective and efficient meta-search engine for document retrieval.

In the next section, we review related works. Section 3 describes data and material used in the research. We present the proposed MSE system in Section 4, together with the whole algorithm description and report results of an experimental study in Section 5. Finally, we conclude the paper in Section 6.

2.Related Work

Personalized ranking search results algorithms for search engines have been proposed to include various types of user information [5] in ranking. To enhance ranking performance and improve search results, algorithms use such information as a user's search context [8], geographical location and searching histories, click-through logs [9], and personal bookmarks [3]. Some algorithms consider the information needs of the user's friends [2, 6, 7]. However, these algorithms largely focus on local activities of the user, and fail to embrace the large social contexts of the user.

3.Data and Material

Facebook is a large popular social networking site. Since the data in this site tends to be rich on social networks, we have chosen it to collect the train and test dataset.

Our MSE obtains web search results from three underlying search (source) engines including Bing, Google and Yahoo.

4. Research Methodology

The search problem is defined as using a set of keywords in a search query, as well as characteristics of the searcher, to identify the most relevant results. To solve the mentioned problem, we propose designing a meta-search engine for document retrieval utilizing social information of the user. In this approach, each user is associated with a user profile that captures his interests available from a social network he or she belongs to. Given a query and the ranked list of documents returned by three certain web search engines, our method first parses the entire list of titles and short descriptions of results and extracts main phrases from them. Then it calculates a vector for each main phrase with parameters like Phrase Frequency/Inverse Document Frequency, Phrase Length, Phrase Independence, Intra and Inter-Cluster Similarity. Then it clusters the vectors by unsupervised learning algorithm. Generated clusters are then ranked on the basis of the user profile. The more similar cluster label to the user’s field of interest gets the higher rank.

4.1 *Clustering Approach*

Clustering methods don't require pre-defined categories as in classification methods. Thus, they are more adaptive for various queries. Nevertheless, clustering methods are more challenging than classification methods because they are conducted in a fully unsupervised way. Organizing web search results into clusters quicken browsing search results. Our method is more suitable for web search result clustering because we emphasize the efficiency of identifying relevant clusters for web users. The clusters are ranked according to their label’s similarity score to the user’s field of interest, thus the more likely clusters required by users are ranked higher.

Self-Organizing Feature Map (SOFM), an unsupervised learning algorithm for clustering problems, is the type of learning algorithms where a system is provided with sample inputs only during the learning phase, but not with the desired outputs. The aim of the system is to organize itself in such a way to find correlation and similarities between data samples [1]. We choose SOFM competitive neural network to cluster main phrase vectors.

4.2 *Phases* of *Proposed* MSE *Algorithm*

1. Receiving web search results from source engines;
   The designed MSE accepts username and query inputs from user and passes the query to three underlying search engines.
2. Removing duplicate results;
   Since meta-search engine receives results from more than one source engine, it has to eliminate duplicates.
3. Extracting main phrases from title and short description of each result;
   Instead of downloading whole page result, the title and short description of each result are parsed in order to filter out stop words and then to extract main phrases. A stop word is a commonly used word such as “the” that most search engines ignore it, both when indexing entries for searching and when retrieving it as the result of a search

query because it waste space and time. Stop words are deemed irrelevant for searching purposes. A main phrase is an n-gram, where $n<=3$, with frequency greater than twice. An n-gram is a subsequence of n words from a sequence of words.

4. Calculating main phrases' vectors;
   We calculate 5 parameters [4] for each main phrase and produce a vector for it.
   1) Phrase Frequency/Inverse Document Frequency (PFIDF)
      This weight is a statistical measure used to evaluate how important a word is to a document. PFIDF property of a main phrase (mph) is defined as:

$$PFIDF(mph) = PF(mph).\log(|D|/DF(mph)) \quad (1)$$

      Where PF(mph) is the total count of main phrase in all the documents D and DF(mph) is the number of documents containing mph. More frequent phrases are more likely to be better candidates of cluster labels; while phrases with higher document frequency might be less informative to represent a distinct label.
   2) Phrase Length (LEN)
      The LEN property is simply the count of words in a main phrase.
   3) Intra-Cluster Similarity of Phrase (Intra-CS)
      If a phrase is a good representation of a single topic, the documents which contain the phrase will be similar to each other. We use Intra-CS parameter to measure the content compactness of documents containing the phrase. First, we convert each document into a vector: d=(x_1,x_2,…). Each component of the vector represents a distinct uni-gram and is weighted by PFIDF of this uni-gram. Intra-CS property of main phrase is the average cosine similarity between its associated documents and its centroid (cen). Intra-CS and cen parameters are defined as:

$$Intra - CS(mph) = 1/PF(mph)\,.\sum_{d\in D\ \&\ mph\in d} cos\big(d, cen(mph)\big) \quad (2)$$

$$cen(mph) = 1/PF(mph)\,.\sum_{d\in D\ \&\ mph\in d} \mathrm{d} \quad (3)$$

   4) Inter-Cluster Similarity of Phrase (Inter-CS)
      Inter-CS property of main phrase is the average cosine similarity between its associated documents and the remainder of the documents. Inter-CS is defined as:

$$Inter - CS(mph) = 1/PF(mph)\,.\sum_{d\in D\ \&\ mph\in d}\ \sum_{d'\in D\ \&\ mph\notin d'} \cos(d, d') \quad (4)$$

   5) Phrase Independence (IND)
      A main phrase is independent when the entropy of its context is high. We confirm independence of main phrase when its left and right contexts are random enough. The followings are the equations for IND and *INDl (or INDr)* which is the independence value for left (or right) context of main phrase, where 0.log0=0.

$$IND(mph) = (INDr(mph) + INDl(mph))/2 \quad (5)$$

$$INDl(mph) = -\sum_{t\in l(mph)} (PF(t)/PF(mph)).\log(PF(t)/PF(mph)) \quad (6)$$

5. Clustering main phrases' vectors by SOFM;
   To provide a suitable representation of input data, produced vectors are normalized. We arrange the neurons of our neural network in a random two-dimensional topology. We use the Link distance function to calculate distances from a particular neuron to its neighbors. The link distance from one neuron is the number of links or steps that must be taken to get to the neuron under consideration.
6. Labeling generated clusters;
   Among the main phrases arranged in a certain cluster, the most frequent one is considered as the label of that cluster. Outlier main phrases are eliminated.
7. Fetching user's field of interest;
   The user's field of interest is retrieved from the prepared data set of user profiles.
8. Computing similarity of each cluster label to the user's field of interest;
   Similarity between two strings is a confidence score that reflects the relation between the meanings of two strings. The similarity is calculated in three steps:
   1) Partitioning each string into a list of tokens (words);
   2) Computing the similarity between tokens using a string edit distance matching algorithm; the string edit distance is the total cost of transforming one string into another using a set of edit rules, each of which has an associated cost. Edit distance is obtained by finding the cheapest way to transform one string into another. Transformations are the one-step operations of insertion, deletion and substitution. In the simplest version substitutions cost about two units except when the source and target are identical, in which case the cost is zero. Insertions and deletions costs half that of substitutions.
   3) Computing the similarity between two token lists; we capture the similarity between two strings by computing the similarity of those two token lists, which is reduced to the bipartite graph matching problem [10]. Given a graph G(V,E), G can be partitioned into two sets of disjoint nodes X(left) and Y (right) such that every edge connects a node in X with a node in Y. X is the set of the first list of tokens. Y is the set of the second list of tokens. E is a set of edges connecting between each couple of vertex (X ,Y), the weight of each edge which connects an x1 to a y1 is computed in previous step. The task is to find a subset of node-disjoint edges with maximum total weight.
9. Classifying results on the basis of the generated clusters of main phrases;
   The results are assigned to relevant main phrases to form final classes of results. A result can not join more than one class. Hence if a result is source of two main phrases belonging to two separate clusters, the cluster with higher similarity is preferred.
10. Ranking results within a class based on the popularity rank of the results;
    For each result we assign a popularity rank that is the result's rank in its source engine. If the result is obtained from more than one source engine, the popularity rank is average of the result's ranks in its multi-source engines.
11. Ranking classes of results on the basis of similarity parameter;
    The produced classes are ranked on the basis of the descending order of computed similarity in phase 8. At last the sorted classes and their labels are listed for user.

5. Results and Analysis

We use breath-first search strategy to crawl the social network data. We first start with five users randomly chosen with different fields of interest, and obtain their friends information. Then we use these friends as the new centers and fetch the friends from these centers. This process was iterative and stopped until a set of 100 different fields of interest was prepared. We select a query of general terms or entity names from one day's query log from Google search engine for each of 100 fields of interest. We choose 60 queries as train dataset. After extracting main phrases and clustering them by SOFM algorithm our neural network was trained.

We use the remained 40 queries as test dataset. These queries are listed in Table 1. We Initialize 40 user profiles. Initializing a user profile is straightforward. A user profile contains username and his field of interest. To test MSE in case of ambiguous queries we prepare a list of 10 challenging queries. For these queries two different fields of interest and hence two distinct user profiles can be considered. So, another 20 user profiles were initialized. The challenging queries and their corresponding distinct fields of interest are listed in Table 2. We use these 60 user profiles to evaluate our proposed MSE against two other meta-search engines, Ixquick and Seekky. We use two standard metrics to evaluate the three meta-search engines: precision and priority. Given a query, let the set of returned documents be K and K′ set of relevant returned documents. Let k′ be a relevant document and r(k′) be the rank of k′ in the returned list of documents, then the precision (Pre) and priority (Pri) parameters are defined as:

$$Pre = Sizeof(K')/Sizeof(K) \tag{7}$$

$$Pri = \sum_{k\prime \in K\prime} (10 - r(k') + 1) \tag{8}$$

Entity Names				General Terms			
Sony	Nokia	World War2	Moon	Trip	Joke	Map	Teacher
Porsche	Egypt	Hollywood	Rhine	Chat	TV	Health	College
Disney	Canada	Panasonic	Libya	Resume	Game	Sport	Story
Obama	Sun	Himalaya	Earth	Time zone	Music	Planet	Pain
Gandhi	Niagara	Persepolis	Dell	Design	Flower	Friend	Analysis

Table 1. 40 queries selected from Google’s query log.

Search ID	Query	Field of Interest
1	Jauguar	Mac OS X 10.2
2	Jauguar	Cars
11	Apple	Trees
12	Apple	Ipods
21	Saturn	Astronomy
22	Saturn	Mythology
31	Jobs	Careers
32	Jobs	Inventors
41	Jordan	Geography
42	Jordan	Fashion
51	Tiger	Golf
52	Tiger	Cats
61	Trec	Research
62	Trec	Environment
71	Ups	Chain Management
72	Ups	Stereography
81	Quotes	Commercial Offer
82	Quotes	Famous Sayings
91	Matrix	Movies
92	Matrix	Hairstyles

Table 2. 10 ambiguous queries and distinct fields of interest.

In case of 40 queries listed in Table 1, the average precision and priority parameters calculated for MSE were almost equal to those of Ixquick and Seekky as they are listed in Table 3. The difference is noticeable in case of 10 challenging queries. The difference confirms the preference of MSE over Ixquick and Seekky as the calculated values are listed in Table 4. The values listed in Table 3 and Table 4 are calculated among first ten results (only the first result page) returned by the three meta-search engines. This way the performance and time-saving properties of three meta-search engines can be comparable. Users rarely go through the next pages returned by a search engine. So the range of precision is 0<=Pre<=1, and the range of priority is 0<=Pri<=55. Higher precision and higher priority values of MSE against Ixquick and Seekky confirm the efficiency and effectiveness of this meta-search engine.

Average Value	MSE	Ixquick	Seekky
Precision	0.838	0.839	0.801
Priority	46.15	46.01	40.99

Table 3. Average values of precision and priority among 40 normal queries.

Average Value	MSE	Ixquick	Seekky
Precision	0.505	0.305	0.28
Priority	24	17.7	16.1

Table 4. Average values of precision and priority among 10 challenging queries.

Given challenging queries, Ixquick and Seekky have problem figuring out the desired information context of user. So they only try to return most popular relevant documents. They work well only in case of a user aiming these popular documents. Using Ixquick or Seekky, user may be forced to go through next pages to get wanted data, whilst using MSE user does not need to go through even the second page and it is one of the benefits of considering user’s field of interest in search process. Figure 2 shows the precision values of MSE in comparison with Ixquick and Seekky per 10 challenging queries of Table 2. Figure 3 shows the priority values of MSE in comparison with Ixquick and Seekky per 10 challenging queries of Table 2. For example consider Search IDs 91 and 92. Ixquick and Seekky return the same result set in case of query, “Matrix” in response to two searchers, one interested in “Movies” and the other interested in “Hairstyles”. But MSE first fetches the user’s field of interest from his profile, clusters the received results and then ranks the generated clusters on the basis of user’s field of interest. All in all the diagram of MSE is placed upper than the other two ones, as it is shown in Figure 2 and Figure 3.

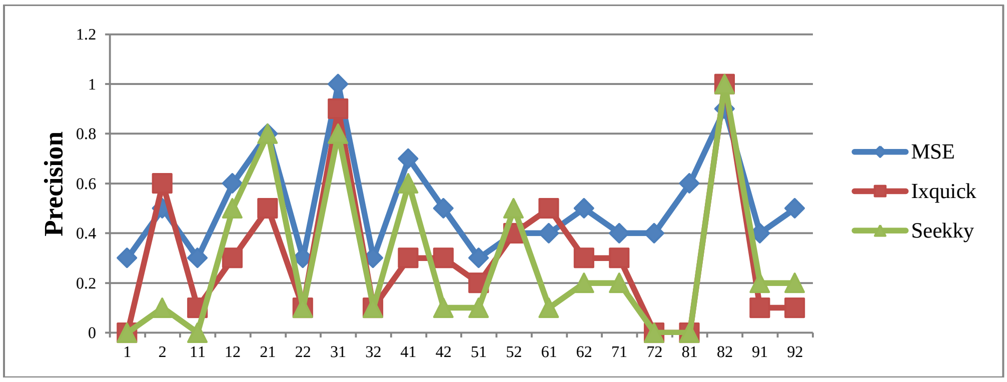


Figure 2: The precision values of three meta-search engines per 10 challenging queries.

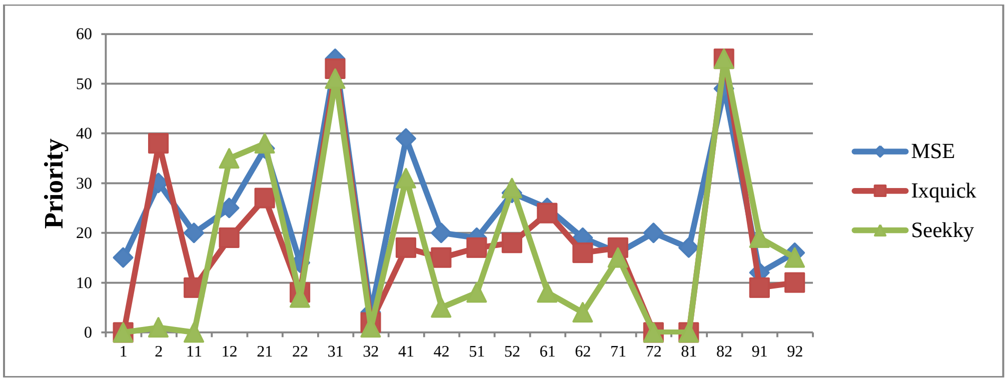


Figure 3: The priority values of three meta-search engines per 10 challenging queries.

6. Conclusions

This research was aimed to test whether the social information of a searcher can improve the ordering of search results on a meta-search engine. In this paper, we present MSE which employs a user-aware approach in ranking the web resources. MSE incorporates both query-document similarity and interest-document similarity to rank search results to fit the searcher's field of interest. Using MSE user receives the required information and search is not a time-consuming task anymore. The proposed method uses a measure of clustering within a set of extracted main phrases, in addition to traditional document relevance, as a feature in matching search queries to possible search results. The results of our experiments show that MSE returns more relevant and higher ranked information in comparison to those ranking methods not considering the user's field of interest.

References

[1] Beale, R. & Jackson, T. (1990). "Neural computing: An introduction". IOP Publishing Ltd. Bristol, UK.

[2] Dalal, M. (2007). "Personalized social & real-time collaborative search". *In Proceedings of the 16th International World Wide Web Conference (WWW '07)*, pp. 1285-1286, May 08-12, Banff, Alberta, Canada.

[3] Jeh, G. & Widom, J. (2003). "Scaling personalized web search". *In Proceedings of the International World Wide Web Conference (WWW '03)*, pp. 271-279.

[4] Kaifeng, X., Li, R., Bao, S., Han, D., & Yu, Y. (2008). "SEM: Mining Spatial Events from the Web". *In Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'08)*, pp. 393-404, May 20-23, Osaka, Japan.

[5] Micarelli, A., Gasparetti, F., Sciarrone, F., & Gauch, S. (2007). "Personalized search on the World Wide Web". *In Lecture Notes in Computer Science. The adaptive web: methods and strategies of web personalization*, Vol. 4321/2007, pp. 195-230.

[6] Mislove, A., Gummadi, K. P. & Druschel, P. (2006). "Exploiting social networks for internet search". *In Proceedings of 5th Workshop on Hot Topics in Networks (HotNets'06).*

[7] Montaner, M., López, B., & de La Rosa, J. L. (2003). "A taxonomy of recommender agents on the internet". *Artificial Intelligence Review*. Vol. 19, No. 4, pp. 285-330.

[8] Shen, X., Tan, B., & Zhai, C. (2005). "Ucair: Capturing and exploiting context for personalized search". *In Proceedings of the Information Retrieval in Context Workshop, SIGIR (IRiX'05).*

[9] Sun, J. T., Zeng, H. J., Liu, H., Lu, Y., & Chen, Z. (2005). "Cubesvd: A novel approach to personalized web search". *In Proceedings of the 14th International World Wide Web Conference (WWW'05)*, pp. 382-390, May 10-14, Chiba, Japan.

[10] Tardos, E. & Kleinberg, J. (2006). "Algorithm Design". Pearson Education, Addison Wesley Publishing, Boston.