Utilizing Lightweight Semantics for Search Context Acquisition in Personalized Search

Tomáš Kramár [*]
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 2, 842 16 Bratislava, Slovakia
kramar@fiit.stuba.sk

Abstract

In order to answer the query, a search engine has to find full-text matches in the background document corpus and then order the documents so that the more relevant results are placed higher in the list. An ideal ranking function should understand user's intent – the goal that is expressed via the query keywords, and order the results such that the results matching user's intent are ranked higher. To understand the user's intent, we need to understand semantics of the queries and the documents. There are various approaches that leverage semantics, but they are heavy-weight, require external knowledge bases and are very hard to implement in a highly dynamic, open-corpus domain, such as the Web. In our work, we focus on the omnipresent lightweight semantics coming from the search result documents. We propose a flexible metadata-based context model and propose methods that scope it to short-term interests or expand it with additional data. We identify several sources of contextual data for this model: temporal context in form of behavioral search patterns, activity-based context in form of past queries and social context in form of user similarity.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering, Selection Process, Query formulation*; H.3.4 [**Information Storage and Retrieval**]: Systems and software—*User profiles and alert services*; I.7.5 [**Computing methodologies**]: Document and text processing—*Document analysis*

Keywords

search, personalized search, search context, social networks, user groups, search sessions, behavioral patterns

1. Introduction

A search engine that does not understand the meaning of the query treats all documents the same, looks for the textual matches and orders the resulting documents by a ranking function. As a consequence, the search results contain (often intentionally [1]) mixed set of documents, covering all possible meanings of the query words and the disambiguation is left to the user alone. Search engines work like databases: they crawl and index documents and respond to queries with a list of results. The order of documents depends on the adopted ranking function; a popular ranking function PageRank [19] scores documents by analyzing inbound links: the more links to a document, the more likely it is to appear at the top positions.

This ordering is however not always compatible with user's information needs: a programmer searching for cucumber probably does not want to make a salad (the most popular meaning of the word refers to vegetable), but to use a testing tool with the same name. However, the same programmer at home might as well be looking to prepare a cucumber salad. In modern search engines the ranking function is more robust and considers many features in order to rank the document [1] and PageRank score represents only one document feature that influences the final ranking. Nevertheless, even combining many features and sub-rankers does not guarantee a satisfactory document ranking and the query problems are still present (ambiguity, shortness and clarity).

To face these problems an area of Personalized Search research has been established. Several approaches have been researched, each with the ultimate aim to help the user find the relevant content, without trying to change how humans think, or work. The main idea behind search personalization is to adapt the search results towards the particular user who issued the query. The ranking should reflect user's interests and especially the immediate need that resulted into the query. Each query is issued with a specific intent - the goal that the user wants to fulfill.

Query represents a projection of that underlying goal into the common interface between human mind and the search engine. That projection is often malformed by the mismatch between complexity of our thoughts and limiting

[*] Recommended by thesis supervisor: Prof. Mária Bieliková

[1] List of features used in Microsoft's Learning to Rank Challenge, `http://research.microsoft.com/en-us/projects/mslr/feature.aspx`

simplicity of textual keywords, leading to ambiguous, short and inaccurate queries. The goal of search personalization is to reconstruct as much of the underlying user intent as possible and use that information to better rank the results. This underlying intent is often referred to as search context in the literature (e.g. [15, 23]).

The terminology is different than in the domain of user modeling, where the term context traditionally refers to the attributes of the environment (i.e. user's location, time, her mood, etc.). In the domain of personalized search, the term context is commonly used to describe user's interests and goals at the time of the query. Unlike attributes of the environment, the search context – the goals and interests – are hard to identify, mostly due to the sparsity of interactions with the search engine and the limited form of interaction that takes place during the search. The challenge in capturing the search context lies in extracting as much knowledge from the available data as possible, and possibly finding ways to extend it and enrich it from other sources.

A good source of the underlying intent are the search results: each result that the user clicks reveals a bit more about the interests, yet this information is hidden inside the free-form text that lacks any unified, machine-parsable structure. Search results are often used to discover categories of interests, or to find similarities between users, but the hidden semantics is often left unused.

The problem of semantics in search is actively being worked on in the area of Semantic Search. The strongest selling point of Semantic Search is the ability to write queries in natural language and freely ask questions without translating them into keywords, which leads into more accurate queries. This, however, comes at a cost of building an overlay layer over the searched domain, annotating the existing documents with semantics. Although manageable in closed domains, in an open-corpus domain such as Web, maintaining the semantic annotations is extremely difficult, if not impossible, even when using automated methods, which pushes back the practical mass deployment of Semantic Search.

Recent research in other areas (e.g., [21]) shows that the semantics can be acquired more easily in form of document keywords, although at the cost of reduced quality and power. Full-fledged semantic annotations contain rich data about the underlying document entities, including types and links to connected semantic information. It is this richness and connection that makes it so useful and that allows complex understanding of the document and information inference. On the other hand, these document keywords are only lightweight, they are not linked, nor do they have any attributes beside the textual label. Their main advantage is that they are freely available in any text and even without additional attributes, they still relay some semantic, because they describe the underlying document. Based on this promising research, we formulate our first goal:

Goal 1: Thin the gap between Semantic Search and Personalized Search in open-corpus domain. Propose a context model that can utilize the lightweight semantics for search personalization.

Another issue with search personalization methods is that they focus mostly on the long-term interests that comprise search context. The user model is built as rich as possible, usually from all actions the user has ever made and that are available in some form. This is partly correct, because there must be enough data about the user in order to make confident decisions, but this may also sometimes cause problems. We do not live in an ideal world and users are often interrupted in their normal workflow with various side-tasks that go beyond their usual interests. The problem is when those small tasks and goals overlap with user's long-term profile, yet are completely different. For illustration, a biologist may be interested in the wild jaguar animals and the personalized search system can leverage this kind of information to favor animal-related documents in her day-to-day searching. But if the user is temporarily interested in buying a car, the search for jaguar is suddenly not so useful. We can alleviate this problem by distinguishing between user's long-term interests and her short-term goals, and combining both to fetch the most helpful documents that focus on the intermediate needs. That does not mean that the long-term interests should be abandoned, we just need to find a better way to distinguish when is the right time to use them. This leads to our second goal:

Goal 2: Analyze possible sources of search context for the purposes of acquisition of immediate goals of the user. Devise methods that extend this model with additional information, utilizing the inherent lightweight semantics.

2. Related work

In general, there are two approaches to search personalization:

- Query refinment, which is also commonly referred to as query expansion or query reformulation. The basic idea is that the original query is altered to better express user's intent. The individual terms in a query may be altered, new terms may be added or existing terms may be removed. From one point of view, this is a strategy that ensures that when two users enter the same query, it is changed so that each user runs a different query. This family of approaches is often used to suggest similar queries [4], offer spelling corrections [7], or to solve the index term synonymy problem [10].
- Results reordering is a less transparent strategy. If two users issue the same query, the same query is run, but the personalization occurs at the ranking stage. The user's needs may be taken into account and some documents with content. that is more closely matching user's intent may be ranked higher. Similarly, the documents that are predicted to contain no useful information may be ranked lower. A good example of a result reordering method is the family of learning to rank approaches [8], which employ machine learning to learn user preferences for various search result attributes.

Basically, both strategies achieve the same result – i.e., a personalized list of documents that are likely to interest the user issuing the query. The main difference is in the process itself – the query refinement methods try to personalize the query itself, while counting on the search engine to provide relevant results. The basic premise is that a personalized query leads to personalized search results, which is generally true. Notwithstanding, there may be some queries that are hard, or even impossible to personalize simply by changing the text itself. On the other hand, result processing methods try to personalize the list of search results.

Other perspective that may be used to look at the personalized search methods is the source of the data that it uses to make the decisions. From this point of view, we distinguish these approaches:

- Content based personalized search methods are concerned with the content of the documents which the user is browsing. Similarly to contentbased recommender systems, content-based search personalization methods try to prefer documents that are similar to documents that the user has seen in the past. These methods build the user model from the content of the documents, which they analyze and extract features from (either the full text, or only some kind of metadata). This type of search systems is not limited to documents only, but works with other type of content data, such as texts of the queries. In fact, there are several variations on the sources of the content, such as full text of the documents [16], snippets from the search results [6], document metadata [4] and queries [2].
- Collaborative based personalized search methods that concentrate on similarity between the users [5], instead of similarity between the documents. When a user enters the query it is not the content of the documents, or texts of previous queries that are used to generate the results, but it is the similarity to other users. The contents of the documents or the queries are still used, but only to find users that have similar interests. The search results are then generated with this similarity in mind. The basic assumption of this approach is that if similar users agreed on the relevant results in the past, they will also agree on relevant results in current search.

3. Context model based on lightweight semantics

Although the idea of search context is recurring in the literature, the definitions vary from work to work. We define the search context as the evidence of user's current goal that she is trying to fulfill by issuing the particular query. The search context is not unique for each query – many search queries can share the same goal, in fact, many queries do share the same goal and thus the same search context, because in the process of search, the query is often refined to reflect new knowledge about the problem. Search context is comprised of several independent parts:

- Long-term user preferences and interests that drive most subsequent information needs.
- Short-term needs and goals that provoked the current search. They often stem from long-term interests, but are often mutually orthogonal, so they need to be kept separatly.
- Knowledge-independent attributes of the external environment, such as user's location, mood, or the weather outside. Although these factors arguably impact the expectations the user has about the search results, it happens only in special cases. In order to keep the focus clear on the knowledge-dependent parts of the context model, this part is ignored.

To model the interests and goals inherent in the knowledge-dependent parts of the search context, we use the ubiquitous lightweight semantics in form of document metadata. This model has several advantages:

- Lightweight semantics is ubiquitous, and unlike the full, heavyweight semantics does not have to be maintained.
- It is flexible, yet provides insight on the underlying search goal. Metadata based context model provides a lightweight semantics that reveals the search intent.
- It is easy to build – keywords can be extracted from the text using standard methods from natural language processing, such as TF.IDF.
- It fits into the keyword model of Web search; queries are composed of keywords and the context model is also composed of keywords, which makes it easy to combine query and context models. It is also easy to index and easy to match existing fulltext indices.

The interests in the search context are modeled using various metadata (keywords, tags, terms etc.) coming from the visited pages. These metadata are not provided explicitly, and need to be automatically extracted from the documents.

The automatically extracted metadata, when put under a close inspection, are not always 100% accurate and representative of the actual key concepts from the document. However, it has been shown [3] that the quality is sufficient for user modeling and that when the model is built from many documents, the more relevant metadata will be more frequent than the less relevant ones.

The context model is based on user interests, automatically acquired from her activity with the search engine [11]. We define it as a hypergraph

$$H := < V, E >$$

with a set of vertices

$$V = A \cup P \cup T$$

where A represents a set of users accessing the pages:

$$A = (a_1, a_2, \cdots, a_k)$$

P represents a set of pages

$$P = (p_1, p_2, \cdots p_l)$$

T represents a set of terms

$$T = (t_1, t_2, \cdots, t_m)$$

E represents a set of edges

$$E = (a, p, t) | a \in A, p \in P, t \in T$$

where

$$P \cap T = \emptyset, A \cap P = \emptyset, A \cap T = \emptyset$$

Using this representation is advantageous, as it allows for good denormalization and allows us to track each of the vertice type independently. It may seem intuitive to merge accesses and pages, but this model allows us to abstract page from access, and if the document represented by the URL (page) changes, we can create new vertex in the graph and connect it respectively.

The proposed context model captures all of the user's activity. As a whole it can be seen as a long-term view of user's interests. It purposefully does not explicitly separate the long-term interests and short-term goals of the search context. However, it is designed in a way that allows to easily scope the model to arbitrary part, either continuous or discontinuous.

Various scopes of this model can be used for search personalization. In the most basic use case, the context model can be used unscoped, as a long-term model of user interests. The way this model is used for search personalization depends on the adopted personalization method. For example, in a simple scenario, the personalization method could use this model to calculate similarity between search results and the model using a cosine similarity.

Another advantage of this model is that several convenient operations can be defined that let us combine multiple scopes at once. We define operations of addition and scalar multiplication. Scalar multiplication affects implicit feedback weights associated with each access a from the hypergraph H. Context addition represents the unification of their respective hypergraphs, in other words, the hypergraphs are merged into new hypergraph.

4. Behavioral patterns as a source of search context

The interests of a person can change in intensity and those changes exhibit some sort of pattern that we can analyze and predict. For example, a person can be highly interested in skiing during winter and in that case, during winter, we can boost ranking for documents that deal with skiing. A good example of class of queries that could benefit from such boosting are transactional queries, e.g. in case of a query in form of a sportswear brands, the skiing equipment manufactured by the particular brand should receive higher ranking than other equipment, because the interest in skiing is peaking at this time. The traces of the idea of these patterns, the seasonality of interests, can be found in various places:

- Psychologists have long ago recognized that people are wearing social masks, personas, the "social face the individual presents to the world" it "reflects the role in life that the individual is playing" [9]. Among many personas an individual can have, two should stand out: the persona related to personal life, and persona related to work life. Separating these two personas and using a separate context model for each of them has the potential to bring a model that is focused, similarly to the lean, short-term model, and yet has more data to allow confident adaptation.
- One of the contexts used in recommender systems is context of seasonality [18]. It is based on the similarity of a seasonal aspect of the recommended item with the current season of the year. Good examples are movies with the Christmas theme – users of the recommender system are much more likely to accept such recommendation on and around Christmas, than they are at other time in the year.

Based on our experiences, we hypothesize that the levels of interests that are propagated via search are unstable and change over time; sometimes increasing, sometimes decreasing, and that these changes form repeating patterns. Intuitively, there are many forms of interest drifts, e.g.:

- periodic drifts in interests that are correlated with the season of the year, e.g. winter sports or summer sport;
- drifts in interests caused by the seasonal appearance of the object the person is interested in, e.g. various seasonal produce or sports and cultural events that repeat periodically;
- drifts in interests related to switching between different tasks. In order for these drifts to be worth considering, the duration of the tasks must be sufficiently long and the tasks must repeat periodically. The most widespread task that matches these criteria is a regular job that most people have. We expect that people are changing interests when they are at work, i.e. people search for conceptually different information when they are working than when they relax.

By maintaining separate models for the searcher we could have a model of interests for the given period and provide more relevant results. A search engine could detect if there is a model available for the given moment and use it to personalize the search results.

Different interest drifts have different periodicity, which may range from hours to years. Moreover, the context model for the given period should include data aggregation computed from the complete time period when the interest was active, not just a single moment in the past.

We have used the AOL logs[2] and built two types of temporal contexts: a workweek/weekend context and a business/leisure context. We have analyzed the behaviour of

[2] AOL logs, `http://zola.di.unipi.it/smalltext/datasets.html`

the users [14] with respect to these two contexts using Davies-Boulding cluster separation score and found that:

- Many users exhibit interest switching patterns and although the distribution of interest switching level across the users in our study was uniform, it has shown that there are users who could benefit from context of seasonality (see Figure 1. We have also shown that not all users behave seasonally as we would have expected intuitively and therefore the context of seasonality should be applied carefully and requires further research.
- There is no correlation between the level of temporal interest switching and number of queries issued during that time.
- Users, who switch interests during weekends are likely to also switch contexts during business hours and leisure time.

We have looked on seasonality from point of view of a Web search, but the idea is applicable to a whole range of other problems as well. Seasonality draws from the patterns in user behavior changes, and those patterns are interesting in general, not only when users are fulfilling their information needs, but also other needs, in communication, or in collaboration. We think that seasonality could be studied from other points of view as well, e.g. to see if there are patterns in communication styles that could be used to improve the collaboration between humans.

5. Past queries as a source of search context

The impact of short-term goals on search personalization has been studied only recently by White et al. [22]. They have shown that creating a model that mixes long-term interests with short-term goals can outperform a long-term model in search personalization accuracy.

Now that we know that incorporating short-term goals into the model works, the question remains, how can we capture them in the model. By the definition of the short-term, we should only build the model from recent queries, with the assumption that the user expressed the short-term goals in some, or all of the recent queries. The problem now reduces to finding those recent queries that can be used to build up the model.

The right way to find the related past queries is to do it automatically, to automatically find the boundaries between the different search intents and between unrelated queries. This is a task that is orthogonal with the well studied task of session segmentation. Session segmentation deals with finding the boundaries between sessions, where the session is usually defined as a continuous interaction with the Web. We aim to find the boundaries between search sessions, which are slightly different than the Web sessions in general. The term search session was never formally defined in the literature and its meaning differs in different works, but we assume that search session is a sequence of search related actions with the single underlying informational intent, similarly to [17].

Thus the goal of search session segmentation is to partition the stream of user queries into segments of queries, where each segment is the search session, i.e., holds the condition that all queries that it contains are related to a single underlying goal.

We propose a method for search session segmentation that aims to alleviate the problems encountered by other existing approaches [13], namely:

- going beyond the queries and considering semantics of the search results,
- expanding the semantics with an external knowledge base, in order to gain more data, in a more accurate form,
- considering the perceived usefulness of the search results in the process of segmentation and
- handling the interrupted sessions and reconnecting them with the queries.

We focus mainly on considering the semantic similarity of the queries and search results. The lexical approach that matches queries that share common parts works well for identifying the obvious similarities between the queries – the reformulations, specifications or generalizations of the query. The lexical approach however fails in cases where the queries are dissimilar. In this case, we match the queries using the metadata of the documents clicked from the search results to get better insight into the purpose of the query by aggregating more data than only the query itself provides. We also evaluate the level of page usefulness for the particular query by collecting and analyzing the implicit feedback indicators that the user provides for each page view. Our approach also considers user interruptions and is able to separate intermingled sessions and reconnect interrupted sessions.

We have run an experiment on a sample of manually labeled queries collected on an internal proxy server. We have shown that considering interruptions can improve the performance of lexical similarity and that we can achieve best results when using the semantic similarity that can outperform both temporal and lexical methods. This experiment has warranted the validity of our approach and confirmed that lightweight semantics and implicit feedback can be used to detect short-term goals.

6. User similarity as a source of search context

Having a short-term source of search context for personalization is important, but so is having a model that captures the long-term interests of the user. The more focused is the model, the less data it contains and the harder it is to build.

Suppose that we have a search context model that we would like to use for personalization. This context model captures user's intent and we would like to answer the questions like: What did other users with the same intent do in this situation? Inspired by the collaborative filtering [20] in the area of recommender systems, we try to leverage the natural trait of every human – the trait of following the crowd.

We propose a method that expands the context model of the user by finding similar users [12]. The important question is how to define similarity of the users and we

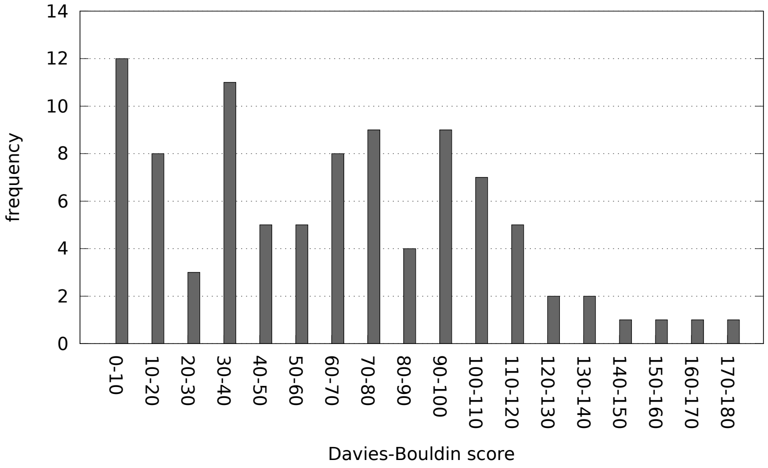


Figure 1: Histogram of Davies-Bouldin cluster scores calculated for the Workweek/Weekday setup assigned to each of the top 100 active users in the AOL logs. Lower values of the score denote more tight and separated clusters.

leverage user's activity and the lightweight semantics inherent in the model to define and calculate this similarity. This method does not depend on the particular context model and can work with any kind of model, as long as it captures the semantics in some way.

We use the proposed context model and extend it with the social context. The social context comes from the similar users, where the similarity is judged by their particular interests. We leverage the metadata based context model, which implicitly captures the interests in form of metadata.

For each user, similar users are found (the community), based on the similarities between their interest-based context models. The process is based on an artificial social-network, where each vertex represents a user and each edge connects two users with a weight that denotes their similarity.

Each user is linked with other users in at least one community. To extend the context model of each user, we simply merge the context model hypergraphs, based on the operation of addition as defined on the context model.

We evaluated our approach on real users, using an internal proxy server to both collect the data for the context models and augment the search results page. The details of the proxy server are outlined in Figure 2.

We designed and evaluated two query expansion strategies, first based on the observation that after an unsuccessful search the query will be reformulated and the second, based on the observation that the keyword meaning can be refined by looking at the document metadata; keywords it frequently co-occurs with. Using a proxy server platform, we integrated our query expansion method into Google search engine and injected our expanded results next to the original results. Figure 3 shows a screenshot of Google with the appended results.

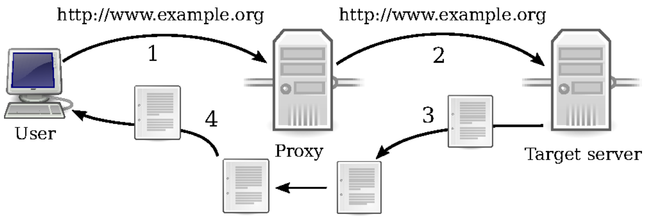
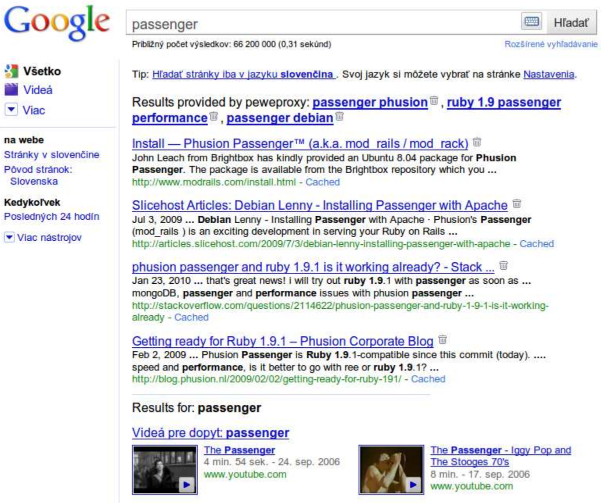


Figure 2: Basic outline of a proxy server and the data collection process. First, the user requests the page (1), the request passes through the proxy, which forwards it (2) to the target server. Its response is sent back to the proxy (3) and then back to the client (4).

We observed that in 70% of the searches where expansion were generated, some of the expanded results were clicked and furthermore, we observed a significant increase in the relevance metrics of the expanded results in comparison with the standard results. A clicked result was considered useful if its dwell time was larger than 4 seconds. In our experiments, only 27% of standard results were considered useful in the baseline environment, while the results expanded by our method were considered useful in 54% of all cases. A post-hoc analysis performed by a human judge revealed that about 63% of all recommended reformulations matched user's intent.

7. Conclusions

The idea of personalizing the search is as old as the search itself and the research area of search personalization is mature and well researched. Nevertheless, the richness of the field only testifies its complexity. The problem of any personalization is hard, but the problem of search personalization is especially tough, because of the mental obstacles that the searcher needs to overcome. First and foremost, a searcher often does not exactly know what she is searching for. Second, users may not have the required vocabulary to correctly describe the problem; even if the goal is clear, the correct query becomes clear only after exploring the problematic area for a while. Third, the queries are almost never exactly aligned with the underlying intent, queries are short and ambiguous. Given all these preconditions, a personalized search must always be one step ahead of the searcher.



Približný počet výsledkov: 66 200 000 (0,31 sekúnd)

Tip: Hľadať stránky iba v jazyku **slovenčina** . Svoj jazyk si môžete vybrať na stránke Nastavenia.

Results provided by peweproxy: **passenger phusion** , **ruby 1.9 passenger performance** , **passenger debian**

Install — Phusion Passenger™ (a.k.a. mod_rails / mod_rack)
John Leach from Brightbox has kindly provided an Ubuntu 8.04 package for **Phusion Passenger**. The package is available from the Brightbox repository which you ...
http://www.modrails.com/install.html - Cached

Slicehost Articles: Debian Lenny - Installing Passenger with Apache
Jul 3, 2009 ... **Debian** Lenny - Installing **Passenger** with Apache · Phusion's **Passenger** (mod_rails) is an exciting development in serving your Ruby on Rails ...
http://articles.slicehost.com/2009/7/3/debian-lenny-installing-passenger-with-apache - Cached

phusion passenger and ruby 1.9.1 is it working already? - Stack ...
Jan 23, 2010 ... that's great news! i will try out **ruby 1.9**.1 with **passenger** as soon as ... mongoDB, **passenger** and **performance** issues with phusion **passenger** ...
http://stackoverflow.com/questions/2114622/phusion-passenger-and-ruby-1-9-1-is-it-working-already - Cached

Getting ready for Ruby 1.9.1 – Phusion Corporate Blog
Feb 2, 2009 ... Phusion **Passenger** is **Ruby 1.9**.1-compatible since this commit (today). .... speed and **performance**, is it better to go with ree or **ruby 1.9**.1? ...
http://blog.phusion.nl/2009/02/02/getting-ready-for-ruby-191/ - Cached

Results for: **passenger**

Videá pre dopyt: **passenger**

The **Passenger**
4 min. 54 sek. - 24. sep. 2006
www.youtube.com

The **Passenger** - Iggy Pop and The Stooges 70's
8 min. - 17. sep. 2006
www.youtube.com

Figure 3: Screenshot of the Google search engine with the experiment in progress. On top, the expanded queries are shown, followed by search results retrieved by query- ing Google with the expanded queries. The injected results are clearly separated from the standard results by a label and a separator line.

Generally speaking, the process of search personalization usually boils down to creating a model which represents user's interests, goals and expectations, which is an inherently complicated task, due to the limited nature of interaction between the searcher and the search engine. Such model can be used to alter the query terms or reorder the list of search results to better suit user's preferences.

Following the goals, we focused on defining search context model and analyzed several sources of search context, each focused on leveraging lightweight semantics. We presented our own contributions, which we now summarize.

Search context model based on document metadata. We have designed a context model that captures the lightweight semantics in form of the ubiquitous document metadata. The main contribution of this model is the flexibility that it offers, it can be both scoped to create new perspectives on the search context or it can be enhanced and extended to accommodate more information. This model supports a limited set of operations that makes it possible to create a linear combinations of various search context perspectives and creation of more powerful context models.

Introducing implicit feedback. We have introduced implicit feedback as measurment of search result relevance and devised a method that can estimate document relevance from various factors of user interaction. The idea of implicit feedback per se is not new, but we are proposing it as an integral part of the search context model. We have shown that the document relevance as perceived by the implicit feedback indicators is an important factor in the search context acquisition, either when searching for similar users, or when finding related searches for the current query.

Perspective on behavioral patterns in search. We have discussed behavioral patterns in search as another possible source of capturing user's goals in the search context model. We have described a study of public log of a search engine and showed that the notion of behavioral patterns is different from what we might think intuitively and that there are users who exhibit behavioral patterns, and there is an equally large group of users who do not.

Method for segmenting search queries into sessions. As part of our focus on short-term goals, we have identified that it is important to know when the user changed her search goal. We have proposed a method that can detect that change and can cluster queries with the same underlying search goals. Main contribution of this method is that it is based on the lightweight semantics that is captured by user's actions and as we have shown with an experiment, can outperform other existing non-semantical approaches.

While this method is completely independent of the proposed search context model, it can be used to scope the context model and offer a perspective on short-term goals.

Method for expanding the search context with data from similar users. Using the document metadata as main signal in the search context is heavily dependent on the metadata quality. Unfortunately, the natural language processing methods as of today are not yet powerful enough to provide 100% relevant metadata. In order to deal with this problem, we propose a model expansion method that utilizes an artificial social network to gather more data. This way, the relevant metadata in the model have a chance of piling up, and subside the lower quality metadata.

Expanding the context model with data from similar users fulfils another important role. It provides a perspective of social context and allows to use the *social mind* as a source of additional knowledge about the underlying query goals. The main contribution of this methods is the way how it can find similar users by analyzing streams of their activity, where the main enabling force is again the lightweight semantics coming from the documents.

Search personalization methods. We have proposed and evaluated two search personalization methods that directly leverage the knowledge contained in the search context model and its graph nature. We have defined *metadata co-occurence analysis* and *query reformulation analysis*. The main contribution of these methods is that they directly leverage the semantics inherent in the context model.

Acknowledgements. This work was partially supported by Scientific Grant Agency of the Ministry of Education of Slovak Republic and the Slovak Academy of Sciences, grant No. VGA 1/0675/11, and Slovak Research and Development Agency, grant No. APVV-0208-10.

References

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA, 2009. ACM.

[2] A. Anagnostopoulos, L. Becchetti, C. Castillo, and A. Gionis. An optimization framework for query recommendation. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 161–170, New York, NY, USA, 2010. ACM.

[3] M. Barla and M. Bieliková. Ordinary web pages as a source for

metadata acquisition for open corpus user modeling. In *Proc. of IADIS WWW/Internet 2010.*, pages 227–233, 2010.

[4] C. Biancalana, A. Micarelli, and C. Squarcella. Nereau: A social approach to query expansion. In *Proceedings of the 10th ACM Workshop on Web Information and Data Management*, WIDM '08, pages 95–102, New York, NY, USA, 2008. ACM.

[5] D. Carmel, N. Zwerdling, I. Guy, S. Ofek-Koifman, N. Har'el, I. Ronen, E. Uziel, S. Yogev, and S. Chernov. Personalized social search based on the user's social network. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1227–1236, New York, NY, USA, 2009. ACM.

[6] P. Ferragina and A. Gulli. A personalized search engine based on web-snippet hierarchical clustering. In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, WWW '05, pages 801–810, New York, NY, USA, 2005. ACM.

[7] J. Gao, X. Li, D. Micol, C. Quirk, and X. Sun. A large scale ranker-based system for search query spelling correction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 358–366, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[8] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 133–142, New York, NY, USA, 2002. ACM.

[9] C. G. Jung. *Two Essays on Analytical Psychology*, volume 7 of *Bollingen XX:7*. Princeton University Press, 1966.

[10] N. Kanhabua and K. Nørvåg. Quest: query expansion using synonyms over time. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III*, ECML PKDD'10, pages 595–598, Berlin, Heidelberg, 2010. Springer-Verlag.

[11] T. Kramár. Towards contextual search: social networks, short contexts and multiple personas. In *Proceedings of the 19th international conference on User modeling, adaption, and personalization*, UMAP'11, pages 434–437, Berlin, Heidelberg, 2011. Springer-Verlag.

[12] T. Kramár, M. Barla, and M. Bieliková. Disambiguating search by leveraging the social network context based on the stream of user's activity. In *UMAP '10: Proc. of the 18th Int. Conf. on User Modeling, Adaptation, and Personalization*, pages 387–392, Hawaii, USA, 2010. Springer.

[13] T. Kramár and M. Bieliková. Detecting search sessions using document metadata and implicit feedback. *Proceedings of the WSCD 2012 Workshop on Web Search Click Data*, 2012.

[14] T. Kramár and M. Bieliková. Analysing temporal dynamics in search intent. In *Proceedings of the 39th international conference on Current Trends in Theory and Practice of Computer Science*, SOFSEM'13, 2013.

[15] S. Lawrence. Context in web search. *IEEE Data Engineering Bulletin*, 23(3):25–32, 2000.

[16] S. Liu, F. Liu, C. Yu, and W. Meng. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 266–272, New York, NY, USA, 2004. ACM.

[17] C. Lucchese, S. Orlando, R. Perego, F. Silvestri, and G. Tolomei. Identifying task-based sessions in search engine query logs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 277–286, New York, NY, USA, 2011. ACM.

[18] L. B. Marinho, I. Nunes, T. Sandholm, C. Nóbrega, J. a. Araújo, and C. E. S. Pires. Improving location recommendations with temporal pattern extraction. In *Proceedings of the 18th Brazilian symposium on Multimedia and the web*, WebMedia '12, pages 293–296, New York, NY, USA, 2012. ACM.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[20] X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009:4:2–4:2, January 2009.

[21] M. Šimko, M. Barla, and M. Bieliková. Lightweight semantics for the "wild web". In *Proceedings of the IADIS international conference on WWW/Internet*, IADIS'11, pages 25–32. IADIS Press, 2011.

[22] R. W. White, P. N. Bennett, and S. T. Dumais. Predicting short-term interests using activity-based search context. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1009–1018, New York, NY, USA, 2010. ACM.

[23] R. W. White and J. Huang. Assessing the scenic route: measuring the value of search trails in web logs. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 587–594, New York, NY, USA, 2010. ACM.

Selected Papers by the Author

T. Kramár, M. Barla, M. Bieliková. Personalizing Search Using Socially Enhanced Interest Model, Built From the Stream of User's Activity. *Journal of Web Engineering*, 12(1-2): 65–92, 2013.

T. Kramár, M. Barla, M. Bieliková. Disambiguating Search by Leveraging a Social Context Based on the Stream of User's Activity. In *Proceedings of UMAP 2010 – 18th international conference on User Modeling, Adaptation, and Personalization*, pages 387–392. Springer-Verlag, 2010.

T. Kramár. Towards Contextual Search: Social Networks, Short Contexts and Multiple Personas. In *Proceedings of UMAP 2011 – 19th international conference on User modeling, adaption, and personalization*, pages 434–437. Springer-Verlag, 2011.

T. Kramár, M. Bieliková. Dynamically Selecting an Appropriate Context Type for Personalisation. In *Proceedings of RecSys 2012 – 6th ACM conference on Recommender systems*, pages 321–324. ACM, 2012

T. Kramár, M. Bieliková. Context of Seasonality in Web Search. In *Proceedings of ECIR 2014 – 36th European Conference on Information Retrieval*. Springer-Verlag. [to appear].

T. Kramár, M. Bieliková. Session Segmentation Based on Document Metadata. *Information Sciences and Technologies. Bulletin of the ACM Slovakia.*, 3(2): 64-66. STU Press, 2011.

T. Kramár, M. Barla, M. Bieliková. PeWeProxy: A Platform for Ubiquitous Personalization of the "Wild" Web. In *Adjunct Proceedings of UMAP 2011 – 19th international conference on User modeling, adaption, and personalization*, pages 7–9, 2011.

T. Kramár, M. Bieliková. Detecting Search Sessions Using Document Metadata and Implicit Feedback. In *WSCD 2012 Workshop on Web Search Click Data 2012*. 2012. [online, accessed Dec. 16 2013] http://research.microsoft.com/en-us/um/people/nickcr/WSCD2012/

T. Kramár, M. Bieliková. Analysing Temporal Dynamics in Search Intent. In *Proceedings of SOFSEM 2013 – 39th Conference on Current Trends in Theory and Practice of Computer Science*, pages 54–63. Institute of Computer Science AS CR, 2013.

T. Kramár, M. Barla, M. Bieliková. Adaptive Proxy Server: Operation and Experience after one Year. (in Slovak) In *Proceedings of WIKT 2010 – 5th Workshop on Intelligent and Knowledge oriented Technologies*, pages 28–31. Institute of Informatics, Slovak Academy of Sciences, 2010.

T. Kramár. Short-term Contexts in Personalized Search. (in Slovak) In *Proceedings of WIKT 2011 – 6th Workshop on Intelligent and Knowledge oriented Technologies*, pages 161–165. Technical University of Košsice, 2011.

T. Kramár, M. Barla, M. Bieliková. User Modeling on the Open Web. (in Slovak) In *Proceedings of Znalosti 2011*, pages 112–123. VŠB – Technical University of Ostrava, 2011.