Context of Seasonality in Web Search

Tomáš Kramár and Mária Bieliková

Faculty of Informatics and Information Technologies
Slovak University of Technology
Ilkovičova 2, 842 16 Bratislava, Slovakia
Bratislava, Slovakia
{name.surname}@stuba.sk

Abstract. In this paper we discuss human behavior in interaction with information available on the Web via search. We consider seasonality as a novel source of context for Web search and discuss the possible impact it could have on search results quality. Seasonality is used in recommender systems as an attribute of the recommended item that might influence its perceived usefulness for particular user. We extend this idea to Web search, introduce a seasonality search context, describe the challenges it brings to Web search and discuss its applicability. We present our analysis of AOL log that shows that the level of seasonal behavior varies.

1 Introduction

It has been recognized that Web search needs some form of information that would help to understand the underlying intent, which is rarely expressed clearly in the query [3]. This information is collectively referred to as a *search context* and there are many sources which the search context can be implicitly inferred from. In this paper we focus on a novel source of search context – context of seasonality. The basic premise behind context of seasonality is that the interests of a person change in intensity and those changes exhibit patterns that we can analyze and predict. E.g., 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.

There are many aspects of this source of context, however, before tackling them deeper, we must first answer the important question whether the basic premise of the seasonality context holds for the Web information space and whether the patterns in user interest shifts exist. It has been shown that seasonality exists at the query level (e.g. a query `ECIR` repeats in a yearly interval) and roughly 7% of all queries are seasonal [8]. According to [1], who analyzed a topically labeled query log, some topics exhibit global popularity peaks throughout the day, while others remain constant. Whether seasonality exists at the level of interests of a single user is still an open question. Intuitively, it seems that seasonality is omnipresent, but in reality this must not be necessarily true and we have to be careful in using it to support tasks on the Web.

In this work, we analyze a search engine log for the existence of various patterns and show that the concept of seasonality in interest drifts is in reality not as straightforward as the intuitive notion. There are users who exhibit clean and predictable interest shifts, but there are also users who do not behave seasonally at all. We also discuss the benefits that the context of seasonality could bring and outline future research directions.

2 Context of Seasonality

There is an important difference between the concept of context in the area of recommender systems and in the area of personalized search. In recommender systems, the context is viewed as a set of external attributes of the environment that impact user's immediate preferences, such as the weather or location and many others. Traditionally, in Web search the context describes any information that can be used to infer the specific goal that the searcher wants to fulfill by issuing a query [6].

Although there are cases when the search context has been established explicitly [11], methods that require less cognitive load by capturing the context implicitly are more preferable. They can be characterized by the source from which they draw the information about user's need. Some of the most intensively studied sources of context in Web search are:

- *Similarity between people*; the underlying intent is inferred from behavior of similar searchers, who are grouped into ad hoc communities [10]. The communities are created based on the chosen similarity criteria (e.g., similar past queries, similar browsing behavior, etc.) and serve as a source of context in the personalization process. When a member of the search community issues a query, the search intent can be clarified by analyzing the preferences of other community members for the particular query [4].
- *User's activity*; where the context is inferred from previously entered queries and behavior on search engine results page [12]. In its simplest form every clicked search result in the predefined time window is incorporated to the context model, which represents a model of searcher's short-term interests. This means that documents matching the short-term interests can be boosted to receive higher ranking.

One of the contexts used in recommender systems is context of seasonality [7]. 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 are much more likely to accept such recommendation on and around Christmas, than they are at other time in the year.

Our idea of seasonality context for search is based on a similar idea. Based on our experiences, we hypothesize that the levels of interests that people have are unstable and change over time; sometimes increase, sometimes decrease, 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 a seasonality context for the searcher we could have a model of interests for the given time and provide more relevant results. A search engine could detect if there is a seasonality context available for the given moment and use it to personalize the search results.

It is important to distinguish the patterns in the interest shifts. Simply looking back to one discrete moment in the history to see which interests were relevant in the past is not enough, because it is not clear which point in the history should we look at. Different interest drifts have different periodicity, which may range from hours (like in the example of work/leisure) to years. There are techniques of time series analysis [13,9] that can be applied to this problem.

Seasonality context could be problematic in situations when the active interest changes unexpectedly. This is not very probable for naturally developed interests, but more probable in situations when the interest was related to a certain task that is now complete and the user no longer has any interest in it, e.g. when a work task is completed and the employee is assigned to a different project, possibly from a different domain. Other weak point is the range of data that must be available in order to discover the repeating patterns. Despite its shortcomings, we believe that context of seasonality would bring benefits into the area of personalized search.

3 Search Engine Log Analysis

In order to answer the basic question – whether the periodic shifts in interests occur, we analyzed the publicly available query log from the AOL search engine[1]. Given that the AOL dataset spans only a period of 3 months, our goal was to find shorter periods of interest drifts and we concentrated on analyzing the existence of the task-related interest drifts related to the searcher's job.

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

We analyzed two different sets of disjoint periods where we expected difference in search intents:

- workdays (Monday-Friday) and weekend (Saturday or Sunday) – *working days* setup;
- working time (9:00-17:00) and leisure time (17:00-9:00), workdays strictly – *working hours* setup. The times were chosen as the typical business hours in the USA.

To characterize and compare search intents we built a topic model for each period using the period's clicked search results. The topic model is a vector of words and leverages lightweight semantics [5] in form of page keywords, page description (both provided by the page authors), title and ODP[2] topics. We extracted these lightweight semantics for each search result clicked in the particular period and added the words from these sources to the topic model of that period.

To compare the topic models of the periods we used Davies-Bouldin score [2], a metric commonly used in evaluating clustering methods. This metric awards clusters with low intra-cluster distances (high internal density) and high inter-cluster distances (well separated clusters). The lower the Davies-Bouldin score, the more tight and well-separated are the clusters indicating more tight and separated interests.

We applied this analysis to the top-100 most active users in the AOL log. Figure 1 shows the distribution of the Davies-Bouldin score over the top-100 AOL searchers in both setups (working days and working hours).

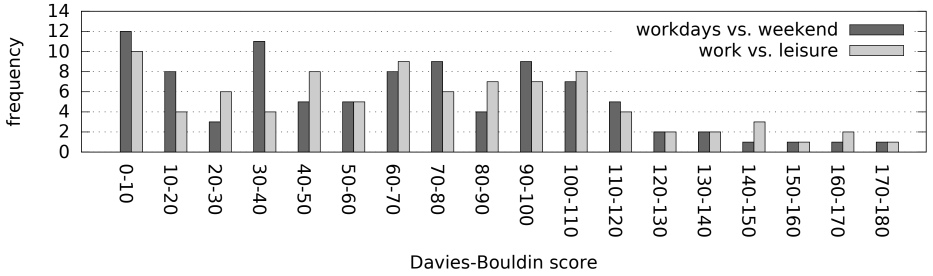

Fig. 1. Distribution of Davies-Bouldin scores across the studied dataset.

We have investigated the correlation between users who exhibit the switching behavior for *working days* setup and users who exhibit the switching behavior for the *working hours* setup. We have ranked each user with the position based on the Davies-Bouldin score, i.e. in each setup, the user with lowest (best) value of Davies-Bouldin score gets ranked with 1, the runner up with 2, continuing this way all the way through the list of users.

[2] Open Directory Project, `http://www.dmoz.org/`

Using the working days setup as a baseline, we have then calculated the change in position for each of the users in working hours setup. The average positional leap in the top-100 dataset is 17 positions. However, if we look only at the top users with best Davies-Bouldin score, the average positional leap of top-5 users is 1 and the average positional leap of top-23 users is 3.17. After the 23rd position, we observe a dramatic increase in the positional leap values. This fact suggests that users who switch interests during working days and weekends are likely to switch interests during working hours and leisure hours, indicating a strict separation of work and free time.

The main discoveries from this log analysis can be summarized in the following points:

- There is no polarization in the interest drifts during the temporal periods in the two selected setups. There are some users who behave according to the intuitive notion of seasonality, i.e., they switch interests during leisure time, but roughly the same amount of users does not exhibit this behavior. In fact, the level of interest switching has uniform distribution. This indicates the need for a further research to find methods to predict the interest switching behavior of the particular user.
- Users, who switch contexts during weekends, are likely to also switch contexts during working hours and leisure time, indicating that a level of switching in one scenario can be used to predict a level of switching in different scenario.

4 Conclusions and Future Work

We have 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.

Introducing the context of seasonality brings many challenges to the area of Web search personalization. First, we need to devise methods to automatically find interest switches patterns and predict their occurrences. These methods would need to operate at the scale of Web search and also need to handle edge cases, like overlapping periods of peaking interest.

One of the benefits of context of seasonality is that it should be applicable in more situations than other sources of context. However, having an ever-applicable source of context raises new questions about composability of search contexts from different sources. Is there a way that contexts coming from different sources could be combined? Or is there a way to compare contexts and always select the better one? These are the questions that should be addressed in further research by the Web search community.

Although we have looked on seasonality from point of view of a Web search, 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, e.g. to see if there are patterns in communication styles that could be used to improve the human collaboration.

Acknowledgements. This work was partially supported by projects No. APVV-0208-10 and No. VG1/0675/11, and it is partial result of the Research and Development Operational Programme for the project University Science Park of STU Bratislava, ITMS 26240220084, co-funded by the European Regional Development Fund.

References

1. Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., Frieder, O.: Hourly analysis of a very large topically categorized web query log. In: Proc. of the 27th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. pp. 321–328. SIGIR '04, ACM, New York, NY, USA (2004)
2. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1(2), 224–227 (1979)
3. Downey, D., Dumais, S., Liebling, D., Horvitz, E.: Understanding the relationship between searchers' queries and information goals. In: Proc. of the 17th ACM Conf. on Information and Knowledge Management. pp. 449–458. CIKM '08, ACM (2008)
4. Kramár, T., Barla, M., Bieliková, M.: Disambiguating search by leveraging a social context based on the stream of user's activity. In: Proc. of the 18th Int. Conf. on User Modeling, Adapt., and Pers. pp. 387–392. UMAP'10, Springer-Verlag (2010)
5. Kříž, J.: Keyword extraction based on implicit feedback. Bulletin of ACM Slovakia 4(2), 43–46 (2012)
6. Lawrence, S.: Context in web search. IEEE Data Eng. Bulletin 23(3), 25–32 (2000)
7. Marinho, L.B., et al.: Improving location recommendations with temporal pattern extraction. In: Proc. of the 18th Brazilian Symposium on Multimedia and the Web. pp. 293–296. WebMedia '12, ACM (2012)
8. Metzler, D., Jones, R., Peng, F., Zhang, R.: Improving search relevance for implicitly temporal queries. In: Proc. of the 32nd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. pp. 700–701. SIGIR '09, ACM, New York, NY, USA (2009)
9. Shokouhi, M.: Detecting seasonal queries by time-series analysis. In: Proc. of the 34th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. pp. 1171–1172. SIGIR '11, ACM, New York, NY, USA (2011)
10. Smyth, B.: Social and personal: communities and collaboration in adaptive web search. In: Proc. of the 1st Int. Conf. on Information Interaction in Context. pp. 3–5. IIiX, ACM (2006)
11. Smyth, B., Coyle, M., Briggs, P.: Heystaks: a real-world deployment of social search. In: Proc. of the sixth ACM Conf. on Recommender Systems. pp. 289–292. RecSys '12, ACM (2012)
12. White, R.W., Bennett, P.N., Dumais, S.T.: Predicting short-term interests using activity-based search context. In: Proc. of the 19th ACM Int. Conf. on Information and Knowledge Management. pp. 1009–1018. CIKM '10, ACM (2010)
13. Zhang, Y., Jansen, B.J., Spink, A.: Time series analysis of a web search engine transaction log. Information Processing and Management 45(2), 230–245 (2009)