WikiPop - Personalized Event Detection System Based on Wikipedia Page View Statistics

Marek Ciglan, Kjetil Nørvåg
Dept. of Computer and Information Science
NTNU, Trondheim, Norway
{marek.ciglan,kjetil.norvag}@idi.ntnu.no

ABSTRACT

In this paper, we describe WikiPop service, a system designed to detect significant increase of popularity of topics related to users' interests. We exploit Wikipedia page view statistics to identify concepts with significant increase of the interest from the public. Daily, there are thousands of articles with increased popularity; thus, a personalization is in order to provide the user only with results related to his/her interest. The WikiPop system allows a user to define a context by stating a set of Wikipedia articles describing topics of interest. The system is then able to search, for the given date, for popular topics related to the user defined context.

Categories and Subject Descriptors: H.4 [Information Systems Applications]: Miscellaneous

General Terms: Algorithms.

Keywords: Wikipedia, recommendation, knowledge base.

1. INTRODUCTION

Wikipedia has received large attention from the computer science research community in recent years. It has been used for solving natural language processing tasks, for enriching information retrieval systems, for ontology building, and for text mining tasks [3]. However, not only the textual content of Wikipedia is of interest. It has also a unique structure, where each article is dedicated to a single topic and articles are densely linked among themselves. In this work, we exploit the the Wikipedia link structure together with its page view statistics in order to make possible a personalized news detection system. By analyzing the trends in the development of page view statistics (number of visits of distinct Wikipedia pages per day), we identify concepts with significantly increased popularity for a given time period. Our hypothesis is that events described in public sources trigger an increase in the number of visits of Wikipedia articles corresponding to the concepts related to those events. Several existing services provide information on which Wikipedia articles received the highest increase in page views for a given date. We, on the other hand, want to identify all articles (concepts) that received a significant increase in page views and present to the user those that are related to his/her interests (user defined context). To identify concepts related to a given context, we use the SA technique over the Wikipedia link graph. The main challenge is how to pre-process the Wikipedia link graph and assign weights to its edges, so that the SA algorithm can retrieve semantically related concepts. The main contribution of this work lies in the proposed edge weighting strategy that combines structural properties of the link graph and correlation of the page views trends between linked articles. As a simple illustrative example, consider a user that is a fan of the performer Michael Jackson, and who defines his context using the Wikipedia concept "Michael Jackson". For the defined context and the date 27.10.2009, the WikiPop system would yield a result set comprising concepts "Michael Jackson's This Is It", "This Is It (Michael Jackson concerts)" and "Kenny Ortega". This would effectively notify him about the upcoming release of the documentary film related to his favorite artist (Figure 1).

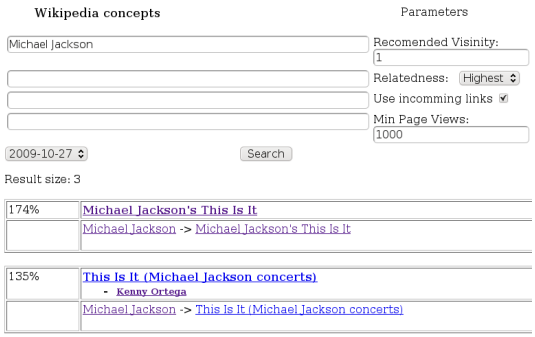

Figure 1: WikiPop prototype GUI

Several existing personalized news detection services, provide personalization based on the similarity between the recommended articles and the articles that the user reads and rates. The novelty of our approach is the utilization of the knowledge base (Wikipedia link graph) to identify the topics that might be of interest to the user based on a defined user context.

2. PERSONALIZED DETECTION OF POPULAR TOPICS

We first studied the possibility of identifying news from Wikipedia page view statistics. Our hypothesis is that events described in public sources, which comprise concepts described by Wikipedia articles, trigger increase in the number of visits of corresponding articles. To verify the hypothesis, we have studied whether we can observe an increase in the popularity of articles related to a set of

events and whether we can identify events, given Wikipedia articles with increased popularity. Although our study was limited in scope due to the requirement of human evaluation, it provided us with an interesting insight. Due to the space limitations, we omit the details of the study. We can, however, summarize that the increase in page views of articles related to real world events is observable from the histogram of page views. In addition, if the popularity increase of a Wikipedia article is also supported by high page views, it is often easy to identify possible source causing increased interest in particular Wikipedia articles (e.g. by using a web search engine, composing the query with the article title and the given date). As the number of article page views gets lower, the success rate drops, even thought the relative increase is high.

In the following, we provide an overview of our approach to the retrieval of topics related to the given context. The context is a set of Wikipedia articles describing topics of user interest. We use the Wikipedia link graph as a knowledge base and the Spreading Activation (SA) algorithm as a method to identify topics related to a context (SA is a method for associative retrieval from a graph data structure [2]). We perform the SA algorithm using concepts from the given context as initially activated nodes. The result of this operation is a set of concepts related to the given context. The final result is the intersection of the set of concepts with increased popularity for a given date and the set of concepts related to defined context (obtained by SA). A straightforward approach, in which the SA is applied on the Wikipedia link graph with constant weights on the edges, produces very large result sets. This is because the Wikipedia link graph is a small-world network with power-law degree distribution and small average distance between nodes. It contains hub nodes with very high indegrees (which usually corresponds to very general topics), through which the activation is spread to a large portion of the network. Thus, we would like to use weights expressing a strength of a semantic relationship between linked concepts. We have proposed a weighting method which has two components. The first function is Indegree Square Ratio (ISR) [1], designed to overcome the problem with the wide spread of the activation through the hub nodes. Let $indegree(i)$ be the indegree of the node i. ISR can be defined as follows: if $indegree(i) > indegree(j)$ then $ISR(e(i,j)) = 1$, otherwise $ISR(e(i,j)) = \frac{indegree(i)^2}{indegree(j)^2}$. The second component of the weighting method is intended to provide the information on the semantic relatedness between topics. It is based on the observation that semantically related articles have similar development trends in page views in the periods of the peak popularity. We argue that this is caused by following reasons: 1) articles targeted by a hyperlink from an article with increased popularity are likely to obtain an increase in page views, and 2) when a topic receives a hight attention, the public tends to look for information on related topics as well. An example is depicted in Figure 2, where we can see the correlation in page view development for the topics a (topic 'Halloween') and b (topic 'Zombie') in the selected period (around the Halloween day), while the third histogram (representing topic Linux) shows different trend in page views development. To measures the relatedness of topics, we define Popularity Development Correlation (PDC) function. For a link $e(i,j)$ between concepts i and j PDC is computed as the Pearson correlation coefficient for the values of page views of i and j for 5 days around the peak popularity of i (PDC is zero iff Pearson correlation < 0). The final weight $W(e(i,j)) = ISR(e(i,j)) \times PDC(e(i,j))$. The following experiment was set up to evaluate the weighting and retrieval of related concepts: from a randomly chosen concept with increased popularity we did a two-hops random walk following the links that can be found in the first paragraph of related Wikipedia articles leading to the node with the same or lower order of magnitude of indegree. From the node where random walk ended, we performed SA algorithm on the link graph with uniform weights on edges and compared it with results of SA algorithm on the link graph weighted by the proposed method. The experiment showed that results obtained on the weighted network were in size 18% of the baseline result set size while in 87% of the cases the original concepts were still present in the reduced result set.

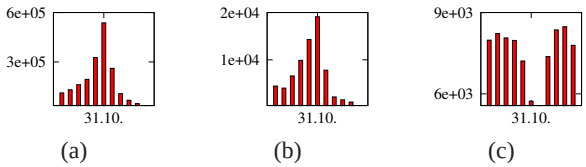

Figure 2: Histograms of page views for Wikipedia articles, period 26.10-05.11.2009: (a) Halloween, (b) Zombie, (c) Linux

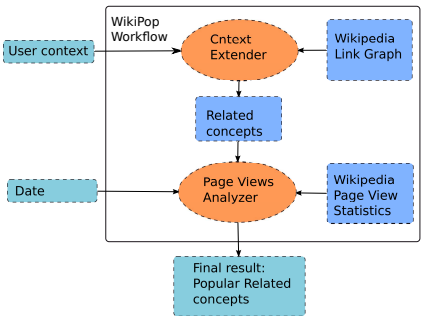

Figure 3: WikiPop workflow

3. DEMONSTRATION SCENARIOS

The WikiPop service prototype is available for demonstration purposes at *http://research.idi.ntnu.no/comidor/wikipop/*. Figure 3 depicts the workflow of WikiPop service. The set of Wikipedia concepts forminga users defined context is extended by the Context Extender mdule (utilizing the SA algorithm over the Wikipedia link graph). The intermediate result is the set of concepts related to the given context. The Page Views Analyzer module checks for increased popularity of concepts in intermediate result for a given date and produces the final result set. WikiPop service is implemented in Java programming language, as a service with a XML-RPC interface. Implementation also includes servelet user interface, so the user can access the service using just a browser. Service uses two data sets - Wikipedia link graph (parsed from Wikipedia XML dump) and Wikipedia page views statistics[1], collected from Wikipedia access log streams and aggregated hourly.

Three demonstration scenarios are planned. In the first scenario, we show correlation between occurring events and the rise in page views for the articles describing concepts related to those events. This will be shown on a set of selected events and corresponding concepts. The second scenario will show operation of the system when a user defined context comprise general topics (e.g. football fan defines his context by concept Football). The third scenario will demonstrate WikiPop results when the user defines context using non-general concepts (e.g. music fan defines context as set of non-main-stream bands).

[1]http://dammit.lt/wikistats/

4. REFERENCES

[1] M. Ciglan, E. Riviere, and K. Nørvåg. *Learning to find interesting connections in wikipedia.* In Proceedings of APWeb, pages 243-249, 2010.

[2] F. Crestani. *Application of spreading activation techniques in information retrieval.* Artif. Intell. Rev., 11(6):453-482, 1997.

[3] O. Medelyan, D. Milne, C. Legg, and I. H. Witten. *Mining meaning from wikipedia.* Int. J. Hum.-Comput. Stud., 67(9), 2009.