Analysis of Topic Dynamics in Web Search

Xuehua Shen
Department of Computer Science
University of Illinois
Urbana, IL 61801
+1 217-244-1036
xshen@cs.uiuc.edu

Susan Dumais
Microsoft Research
One Microsoft Way
Redmond, WA 98052
+1 425-706-8049
sdumais@microsoft.com

Eric Horvitz
Microsoft Research
One Microsoft Way
Redmond, WA 98052
+1 425-706-2127
horvitz@microsoft.com

ABSTRACT

We report on a study of topic dynamics for pages visited by a sample of people using MSN Search. We examine the predictive accuracies of probabilistic models of topic transitions for individuals and groups of users. We explore temporal dynamics by comparing the accuracy of the models for predicting topic transitions at increasingly distant times in the future. Finally, we discuss directions for applying models of search topic dynamics.

Categories and Subject Descriptors

H.3 Information Storage and Retrieval

General Terms

Algorithms, Measurement, Experimentation.

Keywords

Web search, topic analysis, user modeling, topic transition

1. INTRODUCTION

The Web provides opportunities for gathering and analyzing large data sets that reflect users' interactions with web-based services. Analysis of the rich data provided by usage logs promises to lead to insights about user goals, improved quality of search results, and new forms of search personalization. We describe research that examines characteristics of the topics and transitions among topics associated with page visits by users engaged in Web search.

The ability to predict users search and browsing behaviors has been explored by researchers in several areas. The analysis of URL access patterns has been used to improve Web caching [2] [5]. Recent work summarizes the topics that people search for on the Web [1][4][9] and explores how page importance measures like PageRank can be specialized to different topics [3]. Topics are also used to construct user profiles via explicit specification of interests [7] or automatic analysis of Web pages visited [10].

In our experiments, we examine topic dynamics over a 5 week period of time with a large number of users. Instead of inferring topics of interest from queries, which are often very short and ambiguous, we identify topics associated with URLs visited. We report the predictive power of individual models versus models built for groups of users, and consider temporal dynamics.

2. MODELING TOPIC DYNAMICS

To model user's search behaviors we analyze a log of queries, URLs visited, and the topical categories associated with each URL. We analyze the nature and consistency of the topics of URLs a user visits over time. We use the 15 first-level topic categories from the Open Directory Project (ODP), a human-edited directory of the Web [6], (e.g., *Arts, Business*, etc.). Several different models are built to predict the topics based on marginal or Markov transition probabilities, and different user groups. Maximum likelihood techniques with Jelinek-Mercer smoothing are used to estimate the probability distributions.

Marginal models use the overall probability distribution for each of the 15 topics. *Markov models* represent the probabilities of transitioning among topics. The model has 225 states, each representing transitions from one topic to another. *Time-specific Markov models* also estimate the probability of moving from one topic to another, but use different models depending on temporal parameters. *Individual models* use the previous behavior of each individual to predict their current behavior. *Group models* use data from groups of similar individuals to predict the current behavior of an individual. For our experiments, we grouped together individuals who had the same maximally visited topic based on their marginal models. *Population models* use data from the entire population to predict the behavior of an individual.

3. EXPERIMENTS

The basic data consists of a sample of instrumented traffic from MSN Search over a 5 week period from May 22 to June 29, 2004. The data includes Client ID, TimeStamp, Action (Query, URL Visit), and Value (a string for Query, a URL for URL Visit). Our sample consists of more than 87 million actions from 2.7 million unique users. Category tags were automatically assigned to each URL using a combination of direct lookup (for URLs in the ODP directory) and heuristics based on the distribution of categories for the site and sub-sites of a URL (for URLs not in the ODP directory). This technique provided about 50% coverage and assigned an average of 1.11 first-level topics to each URL.

To study topic dynamics, we selected a sample of 6,153 users who had more than 100 actions during the first two weeks. This subset contains more than 660,000 URL visits. These individuals viewed URLs associated with an average of 7.2 different topics. Table 1 summarizes the Markov transition probabilities from one topic to another, using the data from week 1. Rows represent the starting topic and columns the destination topic, and values are normalized by row. Bold numbers show the most common transition in each row. Transitions from a state to itself are most common, but there are some cases where transition probabilities to the most common state (*Arts*) are higher than self transitions.

The main focus of our experiments was to predict the topic of the next URL that an individual will visit over time. The variables we explored were the type of models (Marginal, Markov or Time-Specific Markov), and the cohort group used to estimate the topic probabilities (Individual, Group or Population). We also varied temporal characteristics of the training set. The accuracy of the topic predictions is summarized using the micro-averaged F1 measure, which is widely used in the text classification literature.

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

	Adult	Arts	Busir	Comp	Gam	Healt	Home	Kids_	News	Recre	Refer	Scier	Shop	Socie	Spor
Adult	0.22	**0.24**	0.04	0.11	0.02	0.02	0.01	0.01	0.02	0.05	0.02	0.02	0.08	0.10	0.05
Arts	0.01	**0.49**	0.04	0.07	0.02	0.02	0.02	0.02	0.02	0.04	0.03	0.02	0.09	0.09	0.03
Business	0.00	0.07	**0.32**	0.07	0.01	0.04	0.05	0.01	0.02	0.06	0.04	0.05	0.16	0.07	0.03
Computers	0.00	0.13	0.07	**0.36**	0.03	0.02	0.03	0.02	0.02	0.05	0.04	0.04	0.08	0.08	0.03
Games	0.00	0.13	0.03	0.10	**0.44**	0.01	0.01	0.04	0.01	0.04	0.02	0.02	0.07	0.05	0.02
Health	0.00	0.05	0.05	0.03	0.00	**0.48**	0.04	0.01	0.01	0.03	0.05	0.05	0.07	0.09	0.02
Home	0.00	0.06	0.09	0.05	0.01	0.06	**0.33**	0.02	0.02	0.05	0.03	0.04	0.16	0.07	0.01
Kids&Teens	0.00	**0.15**	0.04	0.07	0.06	0.04	0.04	0.15	0.01	0.06	0.07	0.09	0.07	0.13	0.03
News	0.01	**0.18**	0.08	0.07	0.02	0.05	0.04	0.01	0.13	0.05	0.05	0.04	0.07	0.16	0.05
Recreation	0.00	0.08	0.07	0.06	0.01	0.03	0.03	0.02	0.01	**0.37**	0.03	0.04	0.15	0.06	0.03
Reference	0.00	0.11	0.07	0.06	0.01	0.06	0.03	0.03	0.02	0.04	**0.24**	0.10	0.06	0.14	0.03
Science	0.00	0.07	0.08	0.06	0.01	0.06	0.04	0.04	0.02	0.05	0.10	**0.29**	0.07	0.11	0.02
Shopping	0.00	0.10	0.08	0.05	0.01	0.03	0.05	0.01	0.01	0.08	0.02	0.02	**0.46**	0.05	0.03
Society	0.00	0.12	0.05	0.06	0.01	0.05	0.02	0.02	0.03	0.04	0.06	0.05	0.07	**0.39**	0.02
Sports	0.01	0.12	0.06	0.07	0.02	0.03	0.02	0.02	0.03	0.06	0.04	0.03	0.11	0.07	**0.33**

Table 1. Markov transition probabilities. Week 1, 6153 users.

3.1 Marginal and Markov Models

Figure 1 shows the accuracy for topic predictions using week 1 data for training and week 2 data for evaluation. For the Marginal model, topic predictions are most accurate when using Individual and Group models. The advantage of the Individual and Group models over the Population model shows that users are consistent in the distribution of topics they visit across weeks. Prediction accuracy is consistently higher with the Markov models than with the Marginal models, which shows that knowing the context of the previous topic helps predict the next topic. For the Markov models, topic predictions are most accurate with the Group and Population models. The lower accuracy of the Individual model is due to data sparsity, since many of the topic-topic transitions were not observed in the training period. We explored techniques for smoothing the Individual model with the Group or Population models, but did not find consistent advantages with smoothing.

3.2 Training Size and Temporal Effects

An approach to the data sparseness challenge for the Markov models is to use more data for training. We explored this by using different amounts of training data from weeks 1-4 to predict week 5 data. The predictive accuracy increases as more training data is used. The Individual model shows the largest improvement from 0.301 to 0.347 (15.8%); the Population model improves from 0.379 to 0.385 (1.5%); the Group model improves from 0.381 to 0.409 (7.4%) and shows the highest overall accuracy.

We also explored temporal parameters for the Markov models. We varied the temporal delay between the training (w1-w4) and testing (w5) data. The predictive accuracy increases as the gap between training and testing decreases from 1 month to 1 week. The Individual model improves the most from 0.301 to 0.332 (10.4%); the Population model improves only slightly from 0.379 to 0.381 (< 1%); the Group model improves from 0.381 to 0.398 (4.5%) and shows the highest overall accuracy. We also examined finer-grained dynamics by learning different models for short-term (within session) and long-term (between session) topic transitions. Short-term transitions were defined as successive URL visits that happened within five minutes of each other. Predictive accuracy for the short-term transitions is higher than for the long-term transitions (0.311 vs. 0.301 for the Individual models), reflecting the fact that even individuals whose interactions cover a broad range of topics tend to focus on the same topic over the short term. We believe there is promise in understanding finer-grained temporal transitions, and will continue to explore such models.

4. CONCLUSIONS AND FUTURE WORK

We examined topic dynamics in Web search and developed probabilistic models to predict topic transitions. Group models provide a good balance between predictive accuracy and computational tractability for both marginal and Markov approaches. We considered the influence on model accuracy of temporal proximity and amount of training data. We would like to extend the results with a detailed characterization of the automated tagging process, and with a wider range of techniques for constructing Group models. We believe that better understanding of the dynamics of topic viewing over time will allow us to better personalize search.

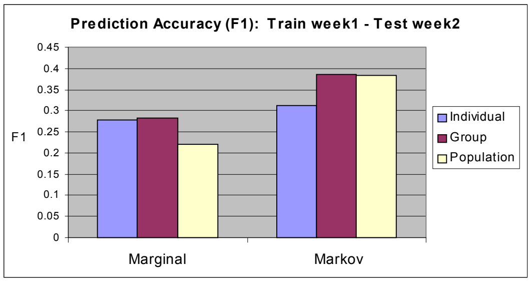

Figure 1. Prediction accuracy (F1) for marginal and Markov models for Individual, Group and Population.

ACKNOWLEDGMENTS AND NOTES

We thank Johnson Apacible, Greg Hullander, Haoyong Zhang, and Robin Wilson for help in data collection. A longer version of this paper is available as a technical report [8].

REFERENCES

[1] Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D. and Frieder, O. (2004). Hourly analysis of a very large topically categorized web query log. *Proceedings of SIGIR 2004*, 321-328.

[2] Deshpanse, M. and Karypis, G. (2004). Selective Markov models for predicting web page access. *ACM Transactions on Internet Technology, 4(2)*, 163-184.

[3] Haveliwala, T. H. (2002). Topic-sensitive PageRank. *Proceedings of WWW 2002,* 517-526.

[4] Lau, T. and Horvitz, E. (1999). Patterns of search: Analyzing and modeling web query refinement. *Proceedings of User Modeling'99*, 119-128.

[5] Lemel, R. and Moran, S. (2003). Predictive caching and prefetching of query results in search engines*. Proceedings of WWW 2003*, 19-28.

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

[7] Ravindran, D. and Gauch, S. (2004). Exploiting hierarchical relationships in conceptual search. *Proceedings of CIKM 2004,* 238-239.

[8] Shen, X., Dumais, S. and Horvitz, E. (2005). Investigations of topic dynamics in Web search. Microsoft Research Technical Report TR-2005-20.

[9] Spink, A., Wolfram, D., Jansen, B. J. and Saracevic, T. (2001). Searching the web: The public and their queries. *JASIST, 52(3)*, 226-234.

[10] Sugiyama, K., Hatano, K. and Yoshikawa M. (2004). Adaptive web search based on user profile constructed without any effort from users. *Proceedings of WWW 2004*, 675-684.