

Ranking Optimization of a Web Page by Means of Latent Models

There are many different approaches to boost the ranking (e.g., the PageRank) of a Web page, for instance, by artificially increasing the number of links from important Web sites that point to this page and by decreasing the number of links that go out of it. Few approaches of ranking optimization regard the content of the Web site itself except for the repetition of terms, which is usually punished by the search engines. Many different scenarios of content manipulation can be thought of that have the purpose of boosting the ranking.

The words in a query do not often match the words in a Web page because of the use of synonyms, and query words that match document words might have different meanings (i.e., polysemy). Several search engines make use of latent topic or semantic models to alleviate these problems. The most known latent model is latent semantic indexing (Deerweester et al. 1990). Probabilistic topic models such as probabilistic Latent Semantic Analysis (pLSA) (Hofmann 1999) and its hierarchical Bayesian form, latent Dirichlet allocation (LDA) (Blei et al. 2003) currently become popular (Steyvers et al. 2007), and model a document as a mixture of topics or aspects. Each topic has its own characteristic word frequency distribution which needs to be estimated from a training set, e.g., using variational inference (EM algorithm) or Gibbs sampling. Knowledge of these models can be used to select the words by which the Web page is built in order to increase the ranking. In addition, these models also open the door to a more refined manipulation of the Web site content. So, one could be interested in boosting a topic, so that it can be retrieved for a query that expresses quite a different topic (e.g., boosting the topic “Latex clothing” for the query “LaTeX”). Or, to put it the other way around, a company might be interested in avoiding certain content words in its Web pages, that would make their ranking quite unpredictable.



The aim of this master thesis is to study the above latent topic models and to design several scenarios of Web page ranking optimization and corresponding manipulation of textual content, to implement a tool for the selection or manipulation of terms in Web pages, to evaluate and compare the rankings based on a retrieval model that incorporates the above topic models. Such a study gives us insights into the capabilities of the latent models to facilitate the manipulation, especially of the probabilistic topic models such as pLSA and LDA. Eventually a Web service can be developed that, for instance, makes suggestions for improvement of the Web text for the ranking towards a particular topic in the assumption that the search engine uses a latent topic model. Software tools for the computation of the above latent topic models and ranking based on these models will be made available by the research group.

We look here for a motivated student with a strong analytical mind and an interest in content spoofing.

Daily supervisors: David De Bock, Wim De Smet, wim.desmet@cs.kuleuven.be, tel.: 016 32 76 02, Celestijnenlaan 200A 3001 Heverlee, Room 03.187.

Promotor: Marie-Francine Moens, sien.moens@cs.kuleuven.be, tel.: 016 32 53 83.

Number of students: 1.

References:

- Deerweester, S. et al. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41 (6), 391-407.
- Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.
- Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of SIGIR* (pp. 50-57). New York: ACM.
- Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D.S. McNamara, S. Dennis and W. Kintsch (Eds.), *The Handbook of Latent Semantic Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.