Toffee - Semantic Media Search Using Topic Modeling and Relevance Feedback

Koho, Mikko

2018-10-11

Koho , M , Heino , E , Oksanen , A & Hyvönen , E A 2018 , ' Toffee - Semantic Media Search Using Topic Modeling and Relevance Feedback ' CEUR Workshop Proceedings , Vuosikerta. 2180 , Sivut 1-4 .

http://hdl.handle.net/10138/299535

Toffee – Semantic Media Search Using Topic Modeling and Relevance Feedback

Mikko Koho[1], Erkki Heino[1], Arttu Oksanen[1,2], and Eero Hyvönen[1,2]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland and
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
http://seco.cs.aalto.fi, http://heldig.fi

1 Research Problem Addressed

This paper considers relevance feedback [1, Ch. 5] search on the Web. Here the information need and query cannot be formulated in the outset—a typical situation in many search situations—but gets refined through making a series of queries and by evaluating the results in between. As an instance of such search the following problem setting is considered: since 1981, The Finnish engineering trade unions TEK and TFiF have given the yearly Finnish Engineering Award[3] to a *"notable engineering or architectural work which has remarkably advanced technical competence in Finland"*. Would it be possible to devise a search system that could help the award committee members in finding out award winning candidates from the news and other materials on the Web?

This paper presents and demonstrates the first results of our research on creating such a search service. The novel idea in the proposed approach is to combine implicit and explicit feedback methods [6] by using topic modeling [2] for extracting topics from the search results. Extracted topics and user feedback are used to generate new search keywords, which then guides the iterative search process. The developed search prototype Toffee is designed to work especially with Finnish language content, but can handle documents in any language.

2 Solution: Topical Relevance Feedback Search

To illustrate the idea, Fig. 1 shows the user interface of Toffee, with a search made to find news related to technology innovations from a web corpus created by the National Broadcasting Company YLE[4]. The initial search was based on the words "innovaatio" (innovation) and "teknologia" (technology). After this, the search has been repeated with feedback that emphasized news articles about the clean technology industry. The actual nine search words are shown below the search field, and the list of results below that. The main topics of each result are shown on the left of the result title and short description. The colored circles indicate different topics and the circle sizes depict the importance of the topic, with a tooltip showing the most important words of a topic.

[3] https://www.tek.fi/en/technology-future/finnish-engineering-award
[4] http://www.yle.fi

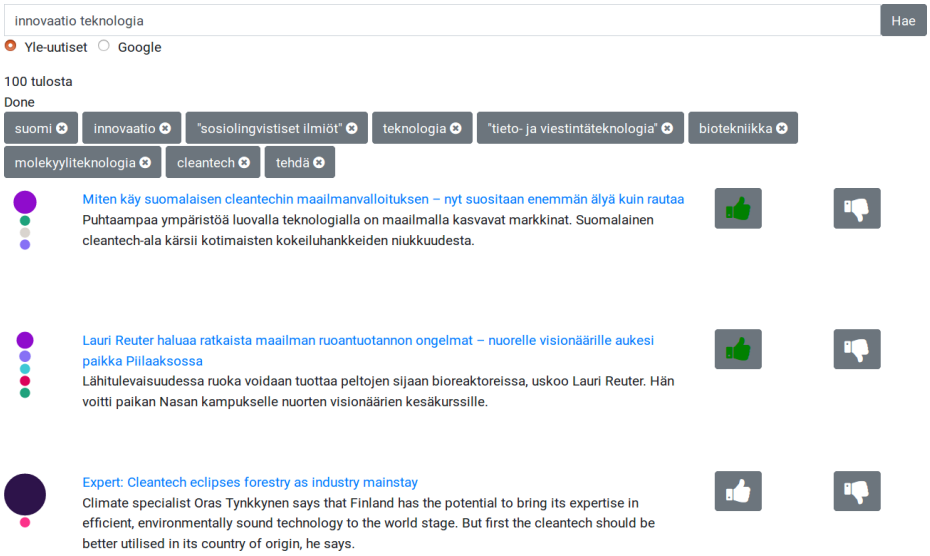

Fig. 1. Toffee user interface, showing the 3 top results of a search.

Toffee source codes are available online[5] as a multi-container Docker application. The search logic is based on the following steps (cf. Fig. 2):

1. A broad initial search is conducted with some keywords that are hypothesized to produce at least some results of interest.
2. Query expansion [11] is applied by looking up broader and similar entities from the Holistic Collaborative Finnish Ontology KOKO [10] via SPARQL using the ARPA annotation tool [3]. The related entities are found by matching entity labels to search keywords and following *SKOS* relations to other entities.
3. The query is sent to the search service API (currently either Google or Elasticsearch) and a maximum of 50 results are received. In case of web search, the resulting web pages are scraped for text contents. With Elasticsearch, the document contents are returned from the search.
4. All the words in the document contents are then reduced to their base forms using the SeCo Lexical Analysis Service [4].
5. Topic modeling is applied to the result set using Latent Dirichlet Allocation [2]. In the case of Elasticsearch, the topics of the whole corpus have been pre-calculated initially. In the case of web search, topics are computed on-the-fly, with a low amount of iterations.
6. All results are returned to the user interface and the user can mark each individual result as interesting or not interesting. The user can then resend the query with the feedback, and the feedback is used to reformulate the query, and the search process continues from step 2.

[5] `https://github.com/SemanticComputing/toffee`

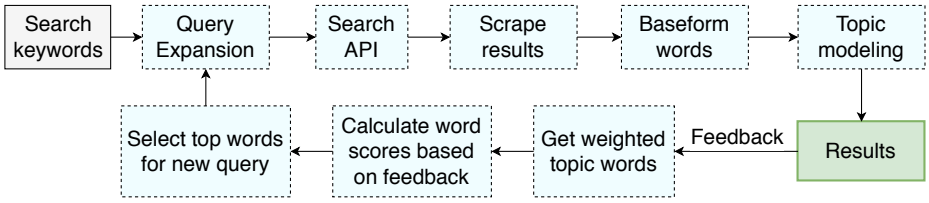

Fig. 2. Iterative Toffee workflow starting from the initial search keywords.

The user can mark any of the returned results as relevant or non-relevant, or leave it undecided. Any of the generated search words of the previous iteration can be removed, as the search process could produce unwanted keywords.

The system reformulates a new iteration of the query, based on the user feedback. An initial weight S_0 is given to each word present in the previous query terms Q or in the words of the previous search results R according to formula 1, where $V = Q \cup R$.

$$S_0(w) = \begin{cases} 1 \text{ if } w \in Q \\ 0 \text{ if } w \notin Q \end{cases}, \forall w \in V \tag{1}$$

The initial weights are then modified based on the possible feedback of each result according to formula 2, where D is the set of result documents from previous search, $\theta_{d,k}$ is the probability of topic k occurring in document d. $\varphi_{k,w}$ is the probability of word w occurring in topic k, and f_d is the user feedback for document d, which can be positive, negative or zero (meaning no feedback is given about the result), with the system using a fixed magnitude to both positive and negative feedback. K is the number of topics.

$$S(w) = S_0(w) + \sum_{d \in D} \sum_{k=1}^{K} \theta_{d,k} \cdot \varphi_{k,w} \cdot f_d \tag{2}$$

The words with the highest weight are then used for the next iteration of the query, with some limit in the maximum number of query words. The user can iteratively give feedback on the results, receive new results, and direct the search to the topics of interest.

3 Related Work and Discussion

Various methods exist for relevance feedback search [1,6]. Teevan et al. [9] enrich web search with relevance feedback based on a constructed user profile. Peltonen et al. [5] combine visual intent modeling with exploratory relevance feedback search. Tang et al. [8] have used topic modeling in academic literature search. Song et al. [7] employed topic modeling with relevance search, based on implicit

feedback from the topics of the user web search history. However, the idea of combining topic modeling of the query results with relevance search, as described in this paper, is to the best of our knowledge new.

Toffee is still in an early stage of development. No formal evaluation about its usability from the user's view point has been made and there are also challenges in measuring precision and recall in an application like this. However, based on our first tests, the idea of providing the user with suggestions for refining the next search seems promising, and if the suggestions seem inappropriate, she is not forced to use them. The system contains plenty of variables to tune, like the number of topics, the number of topic modeling iterations, feedback strength, and query expansion details, which impact the system performance, and the full potential of the approach has not been reached yet. In the future, the system will be evaluated based on the original research problem.

Acknowledgements This research was partially funded by Business Finland and The Media Industry Research Foundation of Finland.

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval (2nd Ed.). Addison-Wesley Longman Publishing Co., Inc. (2011)
2. Blei, D.M.: Probabilistic topic models. Commun. ACM 55(4), 77–84 (Apr 2012), http://doi.acm.org/10.1145/2133806.2133826
3. Mäkelä, E.: Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text. In: Proceedings of the ESWC 2014 demonstration track, Springer-Verlag (May 2014)
4. Mäkelä, E.: LAS: an integrated language analysis tool for multiple languages. The Journal of Open Source Software 1(6) (oct 2016)
5. Peltonen, J., Strahl, J., Floréen, P.: Negative relevance feedback for exploratory search with visual interactive intent modeling. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces. pp. 149–159. ACM (2017)
6. Salton, G., Buckley, C.: Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science 41(4), 288 (1990)
7. Song, W., Zhang, Y., Liu, T., Li, S.: Bridging topic modeling and personalized search. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. pp. 1167–1175. Association for Computational Linguistics (2010)
8. Tang, J., Jin, R., Zhang, J.: A topic modeling approach and its integration into the random walk framework for academic search. In: Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. pp. 1055–1060. IEEE (2008)
9. Teevan, J., Dumais, S.T., Horvitz, E.: Personalizing search via automated analysis of interests and activities. In: Proc. of the 28th Annual International ACM SIGIR Conference. pp. 449–456. SIGIR '05, ACM (2005)
10. Viljanen, K., Tuominen, J., Hyvönen, E.: Ontology libraries for production use: The finnish ontology library service onki. In: European Semantic Web Conference. pp. 781–795. Springer (2009)
11. Voorhees, E.M.: Query expansion using lexical-semantic relations. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 61–69. Springer-Verlag New York, Inc. (1994)