Investigation on Smoothing and Aggregation Methods in Blog Retrieval

Mostafa Keikha
Faculty of Informatics, University of Lugano
Lugano, Switzerland
mostafa.keikha@usi.ch

ABSTRACT

Recently, user generated data is growing rapidly and becoming one of the most important source of information in the web. Blogosphere (the collection of blogs on the web) is one of the main source of information in this category. In my work for my PhD, I mainly focussed on the blog distillation task which is: given a user query find the blogs that are most related to the query topic [3].

There are some properties of blogs that make blog analysis different from usual text analysis. One of these properties is related to the time stamp assigned to each post; it is possible that the topics of a blog change over the time and this can affect blog relevance to the query. Also each post in a blog can have viewer generated comments that can change the relevance of the blog to the query if these are considered as part of the content of the blog. Another property is related to the meaning of the links between blogs which are different than links between websites. Finally, blog distillation is different from traditional ad-hoc search since the retrieval unit is a blog (a collection of posts), instead of a single document. With this view, blog distillation is similar to the task of resource selection in federated search [1].

Researchers have applied different methods from similar problems to blog distillation like ad-hoc search methods, expert search algorithms or methods from resource selection in distributed information retrieval.

Based on our preliminary experiments, I decided to divide the blog distillation problem into two sub-problems. First of all, I want to use mentioned properties of blogs to retrieve the most relevant posts for a given query. This part is very similar to the ad hoc retrieval. After that, I want to *aggregate* relevance of posts in each blog and calculate relevance of the blog. This part requires the development of a cross-modal aggregation model that combines the different blog relevance clues found in the blogosphere.

We use structure based smoothing methods for improving posts retrieval. The idea behind these smoothing methods is to change the score of a document based on the score of its similar or related documents. We model the blogosphere as a single graph that represents relations between posts and terms [2]. The idea is that in accordance with the Clustering Hypothesis, *related documents should have similar scores for the same query.* To model the relatedness between posts, we define a new measure which takes into account both content similarity and temporal distance.

In more recent work, in the aggregation part of the problem, we model each post as evidence about relevance of a blog to the query, and use aggregation methods like Ordered Weighted Averaging operators to combine the evidence. The ordered weighted averaging operator, commonly called OWA operator, was introduced by Yager [4]. OWA provides a parametrized class of mean type aggregation operators, that can generate OR operator (Max), AND operator (Min) and any other aggregation operator between them.

For the next steps, I'm thinking about capturing the temporal properties of the blogs. Bloggers can change their interests over the time or write about different topics periodically. Capturing these changes and using them in the retrieval is one the future woks that I'm interested in. Also, studying the relations between blogs and news and their effect on each other is an interesting problem.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

Blog search, Temporal analysis, User generated data.

1. REFERENCES

[1] J. L. Elsas, J. Arguello, J. Callan, and J. G. Carbonell. Retrieval and feedback models for blog feed search. In *SIGIR*, pages 347–354, 2008.

[2] M. Keikha, M. J. Carman, and F. Crestani. Blog distillation using random walks. In *SIGIR*, pages 638–639, 2009.

[3] I. Ounis, M. De Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the TREC-2006 blog track. In *Proceedings of TREC*, pages 15–27, 2006.

[4] R. Yager. On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Trans. Syst. Man Cybern.*, 18(1):183–190, 1988.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland
ACM 978-1-60558-896-4/10/07.