REPORT ON "A UNIFIED RELEVANCE MODEL FOR OPINION RETRIEVAL"

TEODOR STOEV, YURY BAKANOUSKI

JANUARY 2011

INTRODUCTION

Interest in opinion retrieval is growing because of the increasing popularity of web blogs and forums where people share their personal opinions and reviews on various topics. Opinions not only help other people to make decisions, but also help business and government agencies to collect valuable feedback.

The authors of the paper point out that opinion retrieval differs significantly from classical topic retrieval. First of all, relevant documents should at the same time be relevant to the topic and contain subjective opinions. What is more, text collections consist mostly of informal web data, such as blogs and forums. Thus we are faced with the problem of retrieving and analyzing online opinions precisely and efficiently.

The greatest challenge in opinion retrieval is modeling the information need, which is usually defined by a user query. The user query consists of a small number of keywords – typically content words only, or content words plus cue words. Using such a query directly leads to very low precision, so initial queries need to be enhanced.

Two distinct approaches to opinion retrieval could be identified. The classical approach is the so called two-stage ranking process. In the first stage, documents are ranked by topical relevance only. Then, in the second stage, candidate relevant documents are re-ranked by their opinion scores. These operations result in a high computational overhead.

An alternative approach is the unified ranking process, in which topic retrieval and sentiment classification are combined into a single process. This usually leads to better performance, and may also lead to higher accuracy. In the current paper, this kind of approach is adopted.

SUMMARY

The paper [1] contributes to opinion retrieval in the following three directions:

1. The authors propose a formal framework for opinion retrieval based on the relevance model approach

2. They explore a series of methods to automatically identify the most appropriate opinion words for query expansion
    - query-independent sentiment expansion
    - query-dependent sentiment expansion
    - mixture relevance model
3. Opinion retrieval experiments are performed for the Blog06 and COAE08 text collections

We are going to summarize briefly each of the three points.

THE FORMAL FRAMEWORK

A formal *Opinion Relevance Model* is proposed in the paper [1], which extends the *Relevance Model* of [2]. At first it is assumed that we can obtain an opinion relevance model from a query. The authors use the following notation:

- R : the opinion relevance model for a query
- D : the document model
- V : the vocabulary

Now the *Kullback-Leibler (KL)-divergence* between the probability distributions of the opinion relevance model and the document model is computed:

$$KL(R\|D) = \sum_{w \in V} P(w|R) log \frac{P(w|R)}{P(w|D)}$$
$$= \sum_{w \in V} P(w|R) log P(w|R) - \sum_{w \in V} P(w|R) log P(w|D)$$

Since the sum $\sum_{w \in V} P(w|R) log P(w|R)$ is identical for all documents, the document scores can be computed as follows:

$$Score(D) = \sum_{w \in V} P(w|R) log P(w|D)$$

$P(w|D)$ can be estimated effectively from the interpolation of the maximum likelihood estimation and the occurrence probability in the entire document collection. The authors use Dirichlet smoothing for this purpose:

$$P(w|D) = \frac{freq_D(w) + \mu C_w/|C|}{|D| + \mu}$$

Now the key issue is the estimation of $P(w|R)$, the probability of a word w given the model R. Just like in Lavrenko and Croft's relevance model [2], the probability of words is estimated from a group of documents relevant to the user query.

Since, content words and opinion words contribute differently to opinion retrieval, two separate vocabularies are identified:

- CV : Content word vocabulary
- OV : Opinion word vocabulary

Using these concepts, the opinion relevance model is defined as a unified model for both topic relevance and sentiment. The score of a document is defined as follows:

$$Score(D) = \alpha \sum_{w \in CV} P(w|R) log P(w|D) + \\ + (1-\alpha) \sum_{w \in OV} P(w|R) log P(w|D)$$

The parameter α is introduced to balance the two relevance scores. The CV set can be obtained in well-known ways. The key issue in this paper is how to select the OV set, for which several query expansion techniques have been proposed.

Query Expansion Techniques

The authors have investigated two main approaches to query expansion with opinion words: query-independent and query-dependent. Since both approaches have pros and cons, a *mixture relevance model* is then proposed to combine them.

Query-independent Sentiment Expansion

Three distinct methods for query-independent expansion have been described:

1. *Sentiment expansion based on* ***seed words***

   We simply restrict the OV to some predefined seed words, which are often either strongly positive, or strongly negative sentiment words.

2. *Sentiment expansion based on* ***text corpora***

   In this case, opinion words are obtained from lexical sources. Since there are usually thousands of opinion words in a lexicon, only the most frequent of them are selected to expand the original queries.

3. *Sentiment expansion based on* ***relevance data***

   In this approach, documents are ranked using a function that is learned automatically from training data.

Given a set of query relevance judgments, the *contribution* of an opinion word w is defined as the maximum increase in the mean average precision (MAP) of the expanded queries over a set of original queries, where w is used to expand every original query.

Intuitively, the contribution of a word w refers to how much w can improve the performance of opinion retrieval. The words with the highest contribution are selected for sentiment expansion.

Query-dependent Sentiment Expansion

This approach is based on the idea that a search target is often associated with some particular opinion words. The dependency between the target and the opinion word is expressed like this:

$$P(w|R) \approx P(w|Q) = P(w|q_1, q_2, \ldots, q_n)$$

$$P(w|q_1, q_2, \ldots, q_n) = \frac{P(w, q_1, q_2, \ldots, q_n)}{P(q_1, q_2, \ldots, q_n)}$$

The essence of the approach is to extract opinion words from a set of user-provided relevant opinionated documents. In case we have no such documents, the approach is still applicable with the use of pseudo-relevance feedback. First, documents are ranked by their query likelihood scores, then the top-ranked documents are selected to comprise the relevant document set C.

The joint probability of an opinion word and the query terms is estimated in the following way:

$$P(w, q_1, q_2, \ldots, q_n) = \sum_{D \in C} P(D) P(w, q_1, q_2, \ldots, q_n|D)$$

$$P(w, q_1, q_2, \ldots, q_n|D) = P(w|D) P(q_1, q_2, \ldots, q_n|D, w)$$

The prior probability $P(D)$ is assumed to be uniform.

Mixture Relevance Model

The mixture relevance model combines the two above-mentioned approaches into a single ranking score. Documents are ranked by the interpolation of the scores assigned by the original query, the query-independent sentiment expansion and the query-dependent sentiment expansion:

$$\begin{aligned} Score(D) = \alpha \sum_{w \in Q} P(w|Q) logP(w|D) + \\ + \beta \sum_{w \in OV_1} P(w|R_1) logP(w|D) + \\ + (1 - \alpha - \beta) \sum_{w \in OV_2} P(w|R_2) logP(w|D) \end{aligned}$$

The empirical parameters α and β give weights to each of the three scores.

Experiments

Test Collections

Two benchmark collections, “Blog06” and “COAE08”, have been used to perform several experiments. The first 50 queries from “Blog06” were used for training, and the rest 100 queries were used for testing. No training queries were available for “COAE08”. Both evaluations aim at locating documents that express an opinion about a given target.

External sentiment lexicons have been used as the source of opinion words for both English and Chinese opinion retrieval. The authors correctly point out that for many words it is still debatable whether they are opinionated or not.

Tests Performed

Tests have been performed to evaluate the performance of query-independent sentiment expansion, query-dependent sentiment expansion, and the mixture model. As a CV they use the set of original query terms, and the document priors are set to be uniform. The results are compared to the so called "baseline" approach, which uses the basic relevance model and the Dirichlet smoothing technique only.

Experiments clearly show that the seed word approach has limited applicability. They also confirm that dictionary-based sentiment methods lead to good performance.

The authors claim that pseudo relevance feedback significantly improves opinion retrieval. This claim is based on the comparison of statistical metrics obtained from the experiments.

Strengths, Weaknesses, Future Work

Strengths

- The paper presents the formal definition of a good alternative to two-stage opinion retrieval.
- The authors have conducted a series of experiments, which prove that their approach works well.
- It is believed that the approach performs both faster and better than two-stage opinion retrieval. The theory and experiments performed more or less prove this statement.

Weaknesses

- There is no factual comparison between the two-stage ranking process and the newly presented unified process. It would have been better if some experiments were performed on both approaches, otherwise the superiority of the unified approach could be questioned.

- The authors claim that their approach is not only efficient for opinion retrieval but also improves on topic-based retrieval. Their claim is based on the fact that the new OR method achieves higher statistical measure scores than the "baseline" method.

  In particular, it is claimed that query-independent sentiment expansion is effective for topic-based retrieval. However, there is neither explanation how it can be applied in such manner, nor why this expansion technique leads to effective results.

- Since the new OP approach relies on interpolation of different scores, results might not always be topic relevant and opinionated at the same time. A highly opinionated document with low topic relevance could be scored higher than a more relevant document with lower opinion score, and vice versa.

- The authors have chosen Dirichlet smoothing as a smoothing technique but there is no explanation of why they have decided to do so. There is a detailed comparison of smoothing methods in [3], according to which Dirichlet smoothing truly has the best performance for title queries, but for long queries Jelinek–Mercer smoothing performs better on average.

- The new approach is more complicated than the two-stage opinion retrieval and therefore requires more efforts and resources to implement.

- The paper and the proposed approach do not address some key unresolved issues of opinion retrieval. These are namely semantic orientation of opinion words and implicit opinions.

  Semantic orientation of some opinion words depends on the context, in which case the proposed approach cannot differentiate between positive, negative and no opinions at all.

  As for implicitly expressed opinions, the approach is practically inapplicable to them.

FUTURE WORK

- The paper mentions taking into consideration of the diversity of feedback documents as a future work possibility. Currently, they are simply selected according to their maximum likelihood scores.

- The authors have identified themselves that they should try to vary the document priors. At this time they are uniform. Studying the layout, structure, and user behavior of blogs and forums may be useful for this purpose.

References

[1] Xuanjing Huang and W. Bruce Croft. A Unified Relevance Model for Opinion Retrieval. CIKM 2009.

[2] V. Lavrenko and W. B. Croft. Relevance-based language models. In SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR conference on Research and Development in Information Retrieval, pages 120–127, 2001.

[3] Chengxiang Zhai and John Lafferty. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. 2004.