Facet-based Opinion Retrieval from Blogs

Olga Vechtomova
Department of Management Sciences
Faculty of Engineering, University of Waterloo
200 University Avenue West, Waterloo, Ontario, Canada, N2L 3G1
ovechtom@engmail.uwaterloo.ca
Tel: +1 519 888 4567 ext. 32675
Fax: +1 519 746 7252

Abstract

The paper presents methods of retrieving blog posts containing opinions about an entity expressed in the query. The methods use a lexicon of subjective words and phrases compiled from manually and automatically developed resources. One of the methods uses the Kullback-Leibler divergence to weight subjective words occurring near query terms in documents, another uses proximity between the occurrences of query terms and subjective words in documents, and the third combines both factors. Methods of structuring queries into facets, facet expansion using Wikipedia, and a facet-based retrieval are also investigated in this work. The methods were evaluated using the TREC 2007 and 2008 Blog track topics, and proved to be highly effective.

Keywords

Opinion retrieval, information retrieval, blogs, sentiment analysis.

1. Introduction

In the recent years, there has been a surge of interest in blogs (Web logs) among the general public. Many people see blogs as an opportunity to communicate to others their daily experiences, views, attitudes and opinions on various topics, such as current events, products, companies and other people. In this way a lot of information reflecting people's personal sentiments on various subjects has been accumulated on the Web. Knowing that opinionated information about many subjects exists on the Web, people may want to make use of it in various situations, for example, when choosing a product or service or making an investment decision. However, finding what others think is not always easy using the general-purpose search engines, which may retrieve many pages containing only factual information, such as sales/shopping sites and technical documentation.

In this paper we propose methods of retrieving blog posts containing opinions about an entity expressed in the query, such as a product, person, event or an abstract concept. The proposed methods consist of four main stages:

- Collection pre-processing. The goal of this stage is to keep only the content-related text in each blog post.
- Query processing. At this stage concepts are identified in the query by utilising Wikipedia, the query elements are grouped into facets, and each facet is expanded with related concepts using Wikipedia.
- Topic-based document retrieval. The goal of this stage is to retrieve as many topically-relevant documents in response to the query as possible.
- Opinion-based re-ranking. The documents retrieved in the previous stage are re-ranked using one of the three proposed opinion-ranking methods.

The proposed opinion ranking methods rely on a lexicon of subjective words and phrases, gathered from a variety of manually and automatically built lexical resources, such as FrameNet (Baker et al., 1998), Levin's (1993) verb classes, Wilson's et al. (2005) list of subjectivity clues, etc. The first method uses the Kullback-Leibler divergence (KLD) (Losee, 1990) to weight subjective words, and factors these weights into the document score, the second method ranks documents based on the proximity of query terms to subjective words and the third uses a combination of the two.

The proposed methods were evaluated on the topics of the opinion finding task of the Blog tracks of 2007 (Macdonald et al., 2007) and 2008 (Ounis et al., 2008) TRECs. The opinion finding task of the Blog track of TREC began in 2006. All Blog tracks have the same document collection "Blogs06" (Macdonald and Ounis, 2006), but different sets of 50 topics. The document collection includes 3.2 million permalink documents (88.8Gb), i.e. blog posts. Each topic consists of the standard

TREC components: title, description and narrative, and describes the entity (e.g., a person, product, event or abstract entity) about which the topic creator wants to find opinions. The objective of the task is to retrieve a ranked list of blog posts, which express opinions about the entity (entities) described in the topic. An example of the topic is given in Figure 1.

```
<num> Number: 935 </num>

<title> mozart </title>

<desc> Description:

Find opinions regarding the composer Wolfgang Amadeus Mozart.

</desc>

<narr> Narrative:

All statements of opinion regarding the composer Wolfgang Amadeus Mozart are relevant.  All statements of opinion regarding music authored by Mozart are relevant. Statements of opinions regarding events, festivals, or publications using Mozart's name are relevant.

</narr>
```

Figure 1. An example of the Blog 2007 topic.

Relevance judgements in the opinion finding task were performed on a 5-point scale:

0 – document is non-relevant;
1 – document is relevant, but contains no opinion on the target entity;
2 – document is relevant and contains negative opinion(s) on the target entity;
3 – document is relevant and contains mixed (both positive and negative) opinions on the target entity;
4 – document is relevant and contains positive opinion(s) on the target entity.

Two types of relevance were defined in the task:

- Topic relevance. A document judged as 0 is non-relevant, while a document judged as any of the other labels above is relevant.
- Opinion relevance. A document judged as either 0 or 1 is non-relevant, while a document judged as 2 or 3 or 4 is relevant.

For each of the two types of relevance, the standard performance measures were calculated: Mean Average Precision (MAP), Precision at 10 retrieved documents (P10), and Precision at R, where R is the number of relevant documents for the given topic (R-prec). In this paper, opinion relevance measures will be denoted with subscript *op* (e.g., MAP_{op}), and topic relevance with subscript *rel* (e.g., MAP_{rel}).

While the general idea of faceted retrieval is not new, the paper proposes a novel method of facet identification in the query, expansion of each facet by means of Wikipedia, and use of facets in opinion retrieval. The proposed methods combine a number of components in opinion-based document ranking: facet distance, facet validation, KLD scores of subjective words, and distance between query terms and subjective words. The paper presents detailed evaluation of each contributing component, and discusses its impact on the effectiveness of opinion retrieval. An analysis of performance by query categories (person, event, product, organisation, media, geography and miscellaneous) is also presented, and shows the extent to which different types of queries benefit from the proposed opinion retrieval methods.

The rest of the paper is organised as follows: in Section 2 we review related work on opinion finding, in Section 3 we describe the proposed methods, Section 4 presents the results of the evaluation, Section 5 contains the discussion and analysis of the proposed methods, and Section 6 concludes the paper and outlines possible future research directions.

2. Related work

The Blog track of TREC has served as a valuable platform for the development and evaluation of opinion retrieval methods. Most of the participants in the Blog opinion finding task adopted a two-stage retrieval model. In the first stage, one of the standard IR methods was applied to retrieve the initial document set, and in the second stage, opinion-finding algorithms were used to re-rank this set. Many methods include collection "cleaning", such as HTML tag removal, filtering of the text found in more than one blog, which is regarded as non-content related text (Lee et al., 2008), removal of non-English blog posts, e.g. (Jia et al., 2008, He et al., 2008), and spam filtering, e.g. (Jia et al., 2008). Query processing most commonly

includes phrase identification and query expansion. For example, Wikipedia was used by (Jia L. et al., 2008 and Yang, 2008), WordNet (Fellbaum, 1998) and MiniPar (Lin, 1998) were used by (Jia L., 2008) to identify phrases in the query. The sources for query expansion terms include Wikipedia, pseudo-relevance feedback on external collections (Jia L. et al., 2008; Yang, 2008), and the internal Blog collection (Lee et al., 2008).

There are two main approaches to opinion-based document ranking: classification approach and a lexicon-based approach. SVM (Support Vector Machines) is commonly used in the classification based approaches, for example, by Jia et al. (2008). In the lexicon-based approaches, subjective lexical units could be learned automatically from the internal collection (He et al., 2008, Lee et al., 2008), from external collections, such as product review sites (Lee et al., 2008; Yang, 2008), or sourced from manually and/or automatically built resources, as done, for instance by Yang (2008). Some methods rely on lexicons developed from several sources, for instance, Lee et al. (2008) use a combination of the lexicon from SentiWordNet (Esuli and Sebastiani, 2006) with a lexicon learned from external product review corpora and through the pseudo-relevance feedback process. Amati et al. (2008) propose a method of generating a dictionary of subjective words and their weighting scheme based on information-theoretic measures.

Here are brief descriptions of some of the opinion-based document ranking algorithms that showed high performance in Blog 08 track: One of the approaches proposed by He et al. (2008a) weights words based on their opinion discriminating ability, and uses a query composed of top-weighted subjective words to calculate document matching scores, which are later combined with the topic-relevance document scores. Their other approach uses OpinionFinder (Wilson et al., 2005) to identify subjective sentences, and calculate document scores based on the proximity of query terms to subjective sentences (He et al., 2008b). Lee et al. (2008) calculate a document score as the sum of opinion scores of subjective words occurring in it, normalised by document length. The opinion scoring method proposed by Yang (2008) integrates a number of factors, such as high-frequency subjective words learned from the training corpora, including Blog 2006 dataset and product reviews, low-frequency words occurring in opinionated documents, and expressions containing personal pronouns, such as "I" and "me". Jia et al. (2008) use SVM to classify sentences as opinionated or non-opinionated, then determine whether the sentences are related to the query based on their position in the document with respect to query terms and phrases, and calculate document scores using a number of approaches, such as the sum of SVM scores of the query-related sentences.

Some work on identifying topical subjectivity, i.e. opinions expressed about a target, has been done outside of the Blog track. For instance, Hurst and Nigam (2004) proposed a method of identifying sentences that are relevant to some topic and express opinion on it. They use a classifier, trained on hand-labelled documents to determine if a document is relevant to a topic, and if the classifier predicts the whole document as topically relevant, they apply the same classifier to predict topical relevance of each sentence. For the sentences predicted topically relevant, they apply sentiment analyser, which relies on a set of heuristic rules and a hand-crafted domain-specific lexicon of subjective words, marked with positive or negative polarity. They evaluated their classification method on a set of messages from online resources such as Usenet and online message boards in a specific domain. Their evaluation results show overall precision of 72%. Yi et al. (2003) proposed a method of extracting positive and negative opinions about specific features of a topic. By feature terms they mean terms that have either a part-of or attribute-of relationships with the given topic or with a known feature of the topic. Their method first determines candidate feature terms based on structural heuristics then narrows the selection using either the mixture language model, or the log-likelihood ratio. A pattern-dependent comparison is then made to a sentiment lexicon gathered from a variety of linguistic resources. The method was evaluated on two domains, digital camera and music review articles, using topic relevance judgements performed by the authors, and achieved precision of 87% and recall of 56%.

Much research has been directed towards document classification by sentiment polarity (Dave et al., 2003; Hu and Liu, 2004; Pang et al., 2002; Turney, 2002). The focus of these works is on classifying reviews as either positive, or negative. Pang et al. (2002) evaluated several machine learning algorithms to classify film reviews as either containing positive or negative opinions. Dave et al. (2003) proposed and evaluated a number of algorithms for selecting features for document classification by positive and negative sentiment using machine learning approaches. Turney (2002) developed an unsupervised algorithm for classifying reviews as positive or negative. He proposed to identify whether a phrase in a review has a positive or negative connotation by measuring its mutual information with the words "excellent" and "poor". A review's polarity is predicted from the average semantic orientation (positive or negative) of the phrases it contains. The method, evaluated on 410 reviews from Epinions in four different domains, showed accuracy between 66% and 84% depending on the domain. Hu and Liu (2004) developed a method of identifying frequent features of a specific review item, and finding opinion words from reviews by extracting adjectives most proximate to the terms representing frequent features. The Blog track has a separate polarity task, which in 2007 was defined as a classification task, i.e. the participants had to return unranked sets of documents with positive and negative opinions about the query target. In 2008, the polarity task was an ad-hoc style task, where participants had to return ranked sets of documents with positive and negative opinions on the query target. Overall the performance of track participants in this task was low, which means that this is still an open research problem. Polarity-based retrieval is beyond the scope of this paper, although the methods presented here could serve as the basis for developing polarity-based retrieval algorithms.

3. Methodology

Our approach to retrieving blog posts containing opinions about the concept expressed in the query is a two-stage process. In the first stage, a set of documents is retrieved in response to the query using a topic-relevance ranking method, while in the second stage, this document set is re-ranked using one of the opinion-finding methods described in this section. Any document retrieval model can be used to retrieve the initial document set, and in the evaluation section we report how the proposed opinion re-ranking methods perform with two different first-stage document retrieval methods: BM25 and one of the standard baselines provided by TREC 2008 Blog track organisers. In the following subsections we will describe our approach, starting with the blog collection pre-processing (section 3.1), query processing and expansion using Wikipedia (3.2), the building of a subjective lexicon (3.3) and, in subsequent sections, opinion-based document re-ranking methods.

3.1 Blog collection pre-processing

In all experiments reported in this paper, only the permalink component of the "Blogs06" collection was used. The collection was processed by removing all text enclosed in "<style>" and "<script>" tags, and all non-printing and non-ASCII characters. Tags, such as "
" (line break), "<p>" (paragraph), "<li>" (list element), "<h>", (heading) and "<title>" (webpage title) and corresponding closing tags where applicable, were replaced with newline characters. Other "cleaning" steps included removing the remaining tags, and decoding URI-encoded strings and special characters. Then, each line, in which the number of hyperlinks constituted 50% or more of the total number of words in that line, was removed ("clean50%+" method). The rationale was to remove parts of text, which were the least likely to have opinionated content. We assumed that such lines are more likely to contain advertisements, various website directories or blog navigation links. We compared this method to a more conservative approach ("clean3+" method): a line is removed if it starts with a hyperlink, and if it is one of the 3 or more consecutive lines that start with a hyperlink, however, the "clean50%+" method yielded better results in the document retrieval experiments than "clean3+", as demonstrated in Table 1. The table shows the results of BM25 runs using single terms from the Title section of Blog 06, 07 and 08 opinion track topics. The runs were performed using the BM25 (Spärck Jones et al., 2000) implemented in the Wumpus search system (Büttcher and Clarke, 2005). The BM25 tuning constants b and k_1 were set to 0.1 and 0.75 respectively, as these showed best results in the evaluation on Blog 06 topics. The same b and k_1 values are used in all the runs reported in this paper. In Table 1, "original" refers to the Blog collection, which was used in its original form as distributed by the track organisers, i.e. with all tags and text kept unchanged. Runs marked with * demonstrate statistically significant improvement at .05 level, ** at .02 level (paired t-test) compared to their corresponding runs based on the "original" collection.

Collection pre-processing level	Opinion relevance			Topic relevance		
	Blog 08	Blog 07	Blog 06	Blog 08	Blog 07	Blog 06
original	0.2602	0.2664	0.1918	0.3066	0.3469	0.2828
clean3+	0.3231**	0.3468**	0.2436**	0.3872*	0.4698**	0.3397**
clean50%+	0.3377**	0.3536**	0.2489**	0.4028*	0.4780**	0.3423**

Table 1. Mean Average Precision (MAP) of BM25 "Title" runs with different levels of collection pre-processing.

Since the "clean50%+" collection pre-processing method demonstrated best results, it was used for all subsequent runs reported in this paper. Stemming was not used in our methods, as it proved to deteriorate performance on the Blog 06 dataset.

3.2 Query processing using Wikipedia

In the methods presented in this paper, we used only the Title section of TREC topics, as it most closely resembles queries submitted by users to Web search engines. In fact, topic titles used in Blog 2006 and 2007 tracks are selections of the actual queries submitted by users to a Web search engine, which were later supplemented with traditional TREC-style description and narrative sections by NIST assessors. Blog 2008 topics were created by NIST assessors, following the format of the topics in the earlier years. Our approach to query processing is based on recognition that each information need consists of one or more *facets*. We view a facet as an aspect of the topic about which the user wants to find information. Facet should not be confused with a concept, which is an abstract idea or a representation that stands for instances of a certain entity, such as a physical object, event and person. A concept in turn, can be expressed in a language as a multiword unit, often referred to in IR literature as a *phrase*, e.g., "World Trade Organization", or a single term, e.g. "Cointreau". For example, the title of Topic 861 "Mardi Gras" represents one facet, consisting of one concept, which is lexically represented as a phrase. On the other hand, the title of Topic 1014 "tax break for hybrid automobiles" consists of two facets "tax break" and "hybrid automobiles", each represented by one concept, which in turn is lexically expressed by a phrase. Another example

is the title of Topic 944: "Opera Software" OR "Opera Browser" OR "Opera Mobile" OR "Opera Mini". It may be argued that this topic title contains one facet or theme, i.e. “Opera” software applications, which is represented by four concepts, each expressed by a phrase.

The method described below aims to achieve the following goals by utilising Wikipedia, the online collaborative encyclopaedia:

- identify concepts in the topic titles, by matching them to Wikipedia article titles;
- group them into facets using heuristic rules;
- expand each facet with new concepts by using Wikipedia article redirects and valid abbreviations.

3.2.1 Identifying concepts in the topic titles

The method for identifying concepts by matching them to Wikipedia titles is described below, and is similar to the method used in (MacKinnon and Vechtomova, 2008). Any part of the query that exactly matched a Wikipedia title was treated as a phrase in document retrieval (Section 3.2.3.4). First, we attempted to match the entire query of *n* words, then, if unsuccessful, every subphrase of size *n-1*, then every subphrase of size *n-2* in the unmatched part of the query, and so on until we reached unigrams. The unmatched unigrams were always kept in the query. Stopwords were filtered out only from the single terms in the query. If a phrase that matched a Wikipedia title contained stopwords, they were not removed, such as in “March of the penguins” (Topic 851).

To illustrate the process of identifying concepts in the query, consider the following example: the query “business intelligence resources” (Topic 898) was split into “business intelligence” and “resources”, because Wikipedia has an article with the title “business intelligence”. Similarly, the query “opera software OR opera browser OR opera mobile OR opera mini” (Topic 944) was split into “opera software”, “opera browser”, “opera mobile” and “opera mini”. The Boolean operator OR was removed because it is a stopword.

Inspection of the resulting queries showed that out of 150 topics from Blog 2006, 2007 and 2008 tracks, only the titles of six were incorrectly resolved into concepts. For instance the title “Ford Bell” (Topic 949) was wrongly resolved as two separate concepts: “Ford” and “Bell”, because there was no article about the person Ford Bell in Wikipedia.

3.2.2 Grouping concepts into facets

Whether two concepts represent one facet or two is not easy to determine automatically, and may not always be agreed upon by humans. In an earlier example of Topic 1014, “tax break” and “hybrid automobiles” clearly represent two facets of the topic. The user does not simply want information regarding each of these concepts separately, but information which relates one to the other. According to the topic description, the user wants to “Find opinions on the Federal tax break for purchasers of gasoline-electric hybrid automobiles.” Thus, a relevant document must cover both facets of this topic. On the other hand, in the example of Topic 944 ("Opera Software" OR "Opera Browser" OR "Opera Mobile" OR "Opera Mini"), we consider the four concepts to constitute one facet, and our reasoning is that they are not complementary, which is also evident by the searcher’s use of “OR” operators. In other words, a document may only refer to one of these concepts and be deemed relevant by the searcher.

Ideally, in order to automatically attribute a concept to a facet, we need to determine whether it is close enough in meaning to other concepts in this facet, such that if the user wrote a text describing his/her information need, these concepts could be mutually substituted without the loss of the overall text’s meaning. Linguistically, such “closely” related words could be synonyms, near-synonyms, abbreviations, alternate spelling forms, and in some cases hyponyms, co-hyponyms and hypernyms. While it is possible to implement sophisticated methods relying on machine-readable dictionaries and thesauri, we opted to use a much simpler method at this stage: if two or more concepts in the query have at least one word in common, they are considered to belong to one facet. For instance, out of the 150 topic titles in Blog 2006, 2007 and 2008 tracks, only one topic (944: "Opera Software" OR "Opera Browser" OR "Opera Mobile" OR "Opera Mini") could be considered as containing more than one concept per facet. We decided not to expend effort on implementing a more complex method than this because it is unlikely that the searchers use more than one concept, such as synonyms or alternate spelling forms, to represent one facet. Facets are used in our opinion-based document re-ranking method as will be explained in Section 3.5.

3.2.3 Expanding facets with new concepts

After the concepts and facets were identified in the user’s query, our next goal is to find related concepts for each facet. Our approach is to use Wikipedia page redirects for this purpose. In Wikipedia, some pages are redirecting users to another (target) page, which is usually a more widely used, standard or formal term denoting the same concept. For example, “Winter Olympics” (Topic 933) is redirected to “Winter Olympic Games”. Figure 2 lists all pages that are redirected to it.

Winter Olympics Olympic Winter Games Origins of the Olympic Winter Games Winter olympics Winter Olympic Jeux Olympiques d'hiver Winter Olympic games The Winter Olympics List of Winter Olympics

Figure 2. Pages redirected to "Winter Olympic Games" in Wikipedia.

Wikipedia anchor texts can also serve as potentially useful source of query expansion terms. For example, MacKinnon and Vechtomova (2008) expanded queries in the Complex interactive Question Answering (CiQA) track of TREC, by first mapping the terms and phrases in a query to Wikipedia article titles, and then expanding them with anchor texts pointing to these articles. Elsas et al. (2008) proposed another method of selecting and weighting anchor texts as QE terms, which showed performance gains in the distillation task of the Blog track. While anchor texts may be useful as QE terms in opinion retrieval, we did not explore their use in this paper, focusing instead on the use of page redirects.

3.2.3.1 Limited query expansion

The algorithm for the method consists of the steps presented in Figure 3.

```
1 For each concept (user_concept) that matched a Wikipedia title (as described in Section 3.2.1)
2       Retrieve all Wikipedia pages, the titles of which match user_concept exactly. Rank pages by the
number of times they have been viewed (page_counter)
3       If number of retrieved pages > 1
4               If max(page_counter) > 0
5                       If there is 1 retrieved page with max(page_counter)
6                               If page with the max(page_counter) is a redirecting page
7                                       Add the target page to the facet containing the user_concept
8                               End If
9                       End If
10              End If
11      Else
12              If page is a redirecting page
13                      Add the target page to the facet containing the user_concept
14              End If
15      End If
16 End For
```

Figure 3. Algorithm for limited expansion.

Line 2 of the algorithm requires some clarification. Wikipedia has pages with the same title, which could be due to two reasons:

- Multiple senses of a word or phrase. For example, at the time of this research there were three pages with the titles "Oscar", "OSCAR" and "OScar", the latter being a redirect to "OScar (open source car)".
- Titles in Wikipedia are case-sensitive, and there exist some redirecting pages with the same title written in a different case. For instance there are pages "Brand Manager" and "Brand manager", which both redirect to "Brand management".

Due to these reasons, a concept (called "user_concept" in the algorithm) identified using the method described in 3.2.1, can have an exact match to more than one Wikipedia page titles. If this is the case, we need to select only one page for each user_concept, so that we can use the titles of pages redirecting to this page as query expansion terms. We decided to rely on the popularity of a page as a criterion for selection, i.e. the number of times the page has been viewed. If several pages matched, and two or more pages have equally the highest number of views, i.e. max(page_counter), we deem them ambiguous and discard (line 5). If there is one matching page with max(page_counter) > 0, we check if this page is redirecting to some other page *B*, and if yes, we add the title of page *B* (the target page) to the facet containing the user_concept (line 7). We do the same in cases when only one matching page was found (lines 12-15).

Some examples are: Topic 1027 ("NAFTA"), which was expanded with the title of the target page "North American Free Trade Agreement"; Topic 1045 ("China one child law"), which was split into two facets "China" and "one child law" in the previous stage, and the second facet was expanded in this stage with the phrase "one-child policy". Figure 4 illustrates the limited and full (Section 3.2.3.3.) query expansion processes for the query "NAFTA".

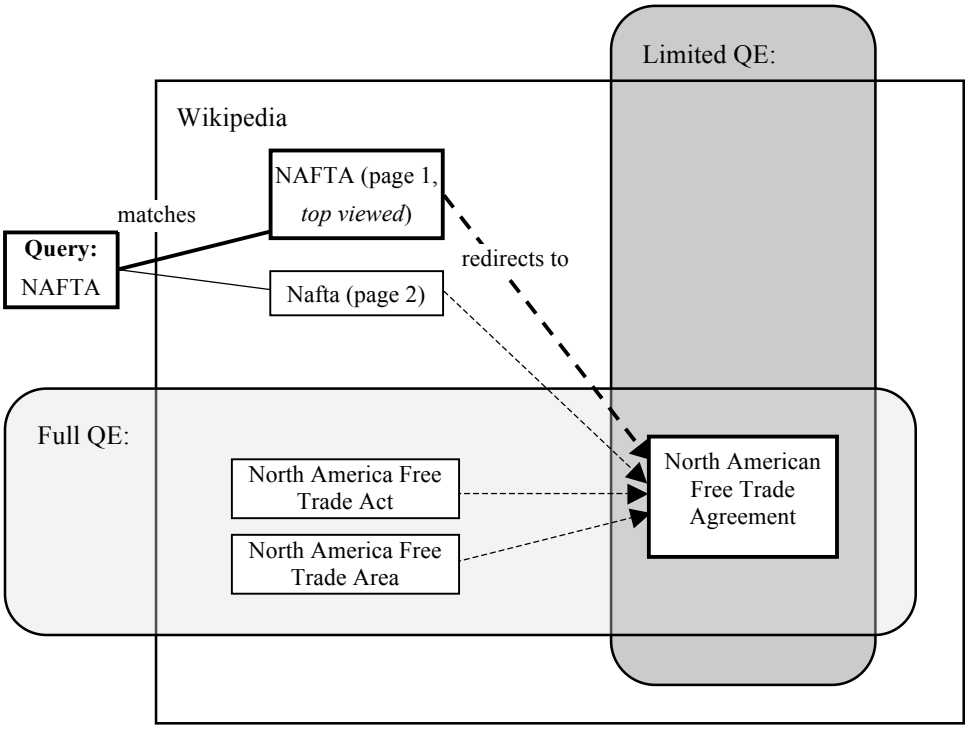

Figure 4. Limited and full query expansion.

3.2.3.2 Expansion with phrase abbreviations

The next query processing step performed was to expand the facets with valid abbreviations of phrases. For each Wikipedia-matched phrase identified so far, either through the initial Wikipedia title matching, or the expansion process, we build an abbreviation, consisting of the initial letters of each of its constituent words. Then, if a page with such a title exists in Wikipedia and redirects to this phrase, it is added to the query facet, containing this phrase. For example, Topic 1013 ("European Union Iceland") was split into two facets in the previous stage: "European Union" and "Iceland". Abbreviation "EU" was among the pages that redirect to "European Union", and was added to the facet containing it. Among the misses of this method was "UN commission on human rights" (Topic 1008), for which the method produced a wrong abbreviation "UCOHR" instead of the correct "UNCHR".

3.2.3.3 Full query expansion

The expansion method described above is rather conservative, and the overall number of expansion concepts found is small. We experimented with adding the titles of all other pages that redirect to the identified target page. In this method, all page titles in Figure 2 redirecting to "Winter Olympic Games", the target page for the query phrase "Winter Olympics", would be added to the facet containing it. Figure 4 shows the pages that would be selected for the query "NAFTA" with limited and full QE methods. The method performed substantially worse than the limited QE method, the reason being likely due to many low-quality and superficially related titles among redirecting pages.

3.2.3.4 QE evaluation

Table 2 presents MAP results of four types of queries: "noQE-single" – the original single terms from topic titles were used; "noQE-Wiki-phrases" – topic titles were matched to Wikipedia page titles as described in Section 3.2.1; "limitedQE-Wiki-phrases" – limited query expansion was performed (Section 3.2.3.1), and valid abbreviations of phrases were added (Section 3.2.3.2); "fullQE-Wiki-phrases" – as previous, plus all other page titles redirecting to the target page were added (Section 3.2.3.3). BM25 implemented in Wumpus was used for all runs. If the query contained phrases, identified using the methods

described in 3.2.1, 3.2.3.1 and 3.2.3.3, then only exactly matching phrases contributed to a document's score, i.e. partially matching phrases were not considered in document scoring. For example, if the query is "Winter Olympics", then only documents containing this exact phrase were retrieved. BM25 score for phrases was calculated by treating the phrase as an indivisible unit, in the same way as a single term. The frequency of the entire phrase in the collection was used to calculate *idf* (inverse document frequency) and the frequency in the current document was used to calculate *tf* (term frequency). In all our methods, weights of query expansion terms were calculated in the same way as for the original query terms. In runs "noQE-Wiki-phrases" and "limitedQE-Wiki-phrases" we also added all single terms from phrases to the query in order to increase recall. In "fullQE-Wiki-phrases" the same single terms as in the previous method were added, but single terms from other redirecting pages were not added, as there are many low-quality titles that are likely to deteriorate performance.

Method	MAP_{op}			MAP_{rel}		
	Blog 08	Blog 07	Blog 06	Blog 08	Blog 07	Blog 06
noQE-single	0.3377	0.3536	0.2489	0.4028	0.478	0.3423
noQE-Wiki-phrases	0.3663*	0.3800*	0.2647*	0.4496**	0.5187**	0.3679**
limitedQE-Wiki-phrases	0.3737*	0.3759	0.2667*	0.4643**	0.5249**	0.3638
fullQE-Wiki-phrases	0.3498	0.3739	0.2476	0.4383	0.5225*	0.3468

Table 2. MAP results of BM25 runs with different levels of query processing

Runs marked with * are statistically significant at .05 significance level, ** at .02 level (paired t-test) compared to their corresponding runs "noQE-single" runs. Mapping topic titles to Wiki phrases ("noQE-Wiki-phrases") improves performance. Limited query expansion method ("limitedQE-Wiki-phrases") is generally better than "noQE-Wiki-phrases", except in Blog-07 MAP_{op} and Blog-06 MAP_{rel}. Full query expansion method is, however, worse than limited QE. Table 3 shows two examples of topics, which were respectively deteriorated and improved with query expansion.

Topic	noQE-single (MAP_{op}; MAP_{rel})	noQE-Wiki-phrases (MAP_{op}; MAP_{rel})	limitedQE-Wiki-phrases (MAP_{op}; MAP_{rel})	fullQE-Wiki-phrases (MAP_{op}; MAP_{rel})
936	"grammys" (0.5173; 0.5463)	"grammys" (0.5173; 0.5463)	"grammys", "grammy award", "award", "grammy" (0.2386; 0.3095)	"grammy", "grammies", "grammy awards", "grammy-winner", "grammy winner", "grammy-award", "the grammys", "grammys", "grammy award", "award" (0.2807; 0.3720)
1039	"geek", "squad" (0.3691; 0.4182)	"the geek squad" (0.4679; 0.6210)	"the geek squad", "geek squad", "squad", "geek" (0.4977; 0.6617)	"geeksquad", "geek squad city", "the geek squad", "geek squad", "squad", "geek" (0.5036; 0.6743)

Table 3. Examples of queries

Since "limitedQE-Wiki-phrases" was generally better than all other query types, we used it as our baseline in the subsequent evaluation of the developed opinion-based re-ranking methods.

3.3 Building the subjective lexicon

The subjective lexicon used in the proposed opinion-based document re-ranking methods was compiled from a number of manually and automatically constructed lexical resources. The manual lexical resources used are Levin's (1993) verb classes, FrameNet (Baker et al., 1998), Ballmer and Brennenstuhl's (1981) speech act verbs, and Hatzivassilouglou and McKeown's (1997) subjective adjectives. We also used a large set of subjectivity clues compiled by Wilson et al. (2006) from both manually and automatically developed lexicons.

3.3.1 Levin's verb classes

Levin (1993) categorised English verbs into semantic classes. Verbs from the following classes were used in our method:

- Verbs of psychological state (e.g., amaze, fascinate, bother, impress);
- Verbs of desire (e.g., crave, yearn, need);
- Judgment verbs (e.g., acclaim, criticize, reproach).

3.3.2 FrameNet

Lexical units (verbs, phrasal verbs, adjectives, etc.) from the following frames were used:

- Emotion_active (e.g., fret, fuss, lose sleep);
- Emotion_directed (e.g., affronted, delighted, resentful);
- Experiencer_object (e.g., enthrall, puzzle, trouble);
- Experiencer_subject (e.g., dissatisfied, jubilant, fed up).

3.3.3 Ballmer and Brennenstuhl (1981) *speech act verbs*

A few verbs and expressions were manually selected from the Emotion Model (e.g., blow up, burst out laughing, grumble about).

3.3.4 Hatzivassilouglou and McKeown's subjective adjectives

A list of 1336 subjective adjectives manually composed by Hatzivassiloglou and McKeown (1997) was used, e.g., amusing, impressive, unreliable.

3.3.5 Wilson's subjectivity clues

Wilson et al. (2006) compiled a collection of subjectivity clues from different sources, including manually developed lexicons, and automatically identified clues from annotated and unannotated corpora. Many of the words in Wilson's et al. collection have been developed using a bootstrapping method reported in (Riloff and Wiebe, 2003). The collection contains 8221 lexical units, including different grammatical forms of some words. Each lexical unit is annotated with the type of subjectivity strength (weakly or strongly subjective), part of speech, and polarity (positive, negative, both, neutral).

3.3.6 Subjective lexicon processing

After the removal of duplicates, the overall subjective lexicon gathered consisted of 6553 lexical units. For each verb and most of the phrasal verbs, whenever it made sense, we also added past tense, gerund ("-ing" form) and third person forms. With all the grammatical word forms added, the total size of the subjective lexicon was 10447.

3.4 Window-based co-occurrence

Our approach to opinion-based reranking consists of adjusting the *tf* weights of query terms on the basis of their co-occurrence with subjective lexical units in fixed-size windows centered around the query term occurrences. The motivation for this is that if a subjective word or phrase occurs close to a query term, it may indicate that the author expresses an opinion about the topic related to the query term. The reason why we chose to use a fixed-size window instead of a natural language unit, such as a sentence, is two-fold: first, a subjective word may not actually "target" the query term occurrence in a sentence, but another word in a different sentence. For instance, a subjective adjective may modify a pronoun in a different sentence, e.g., "He saw an oil painting near the window. It was beautiful." Alternatively, it may modify a noun related to the query term, for instance, when expressing an opinion about a photo camera, a person may talk about the picture quality, rather than the camera directly: "I bought a new camera. The picture quality is excellent." The second reason is practical: it is faster and less error-prone to identify windows than to split the text into sentences. In the case of blogs, sentence boundary detection is a more difficult task than with, say, newswire articles: the former may use non-standard and ill-formed syntactic constructions without proper punctuation marks, and are typically in HTML format, which may not be always possible to convert to plain text correctly.

The window is defined as *n* words to the left and right of the query term occurrence in text. In cases where the distance between two instances of query term(s) in a document is less than *n*, the text span between these two query term instances is split in the middle, such that one half is attributed to the query term on the left, and the other half – to the query term on the right. This is done in order to avoid counting the same subjective word occurrence twice. In our experiments window size was set to 30, as it proved optimal on the training Blog 2006 topics compared to other window sizes (10, 20 and 40).

3.5 Opinion-based document ranking methods

Three opinion-based document re-ranking methods were developed:

- KLD-based method, where document score depends on the Kullback-Leibler divergence scores of the subjective words occurring in the windows around query term occurrences;
- Distance-based method, where distance between a query term occurrence and each of the subjective words co-occurring with it in the window is factored into the document score;
- Combined method, which factors in both distance and the KLD score of each subjective word co-occurring with a query term occurrence in a window.

The methods are described in detail in the following sections. All methods are used in the opinion re-ranking stage according to the algorithm presented in Figure 5. Specifically, all methods contain the facet validation (FV) component, according to which a document is down-ranked if it does not contain at least one concept from each query facet.

```
For each document retrieved in the initial document retrieval stage (baseline)
    Calculate new_document_score using one of the three methods
    Determine the number of query facets found in the document (num_found_facets) out of the total
     number of facets in the query (num_query_facets)
    If num_found_facets < num_query_facets
            Mark document as non_valid_document
    Else
            Mark document as valid
    End If
End If
Rank all valid documents by new_document_score
Append all non_valid_documents to the end of the ranked list in their original rank order in the baseline
run
```

Figure 5. Algorithm for the second stage (opinion re-ranking).

A document is considered to contain a facet if at least one concept (a phrase or single term) from that facet is found in the document. Consider Topic 1045 (Facet 1: "China"; Facet 2: "one child law", "one-child policy"). To be considered "valid" according to the facet validation algorithm, a document must contain "China" and either "one child law", or "one-child policy".

3.5.1 KLD-based method

3.5.1.1 Calculating KLD scores of subjective lexical units

Intuitively, subjective words that occur relatively more frequently in the known relevant and opinionated documents than in the non-relevant or relevant but non-opinionated ones are more useful in predicting opinions. Based on this intuition, we calculated scores for the units in our subjective lexicon using the Kullback-Leibler divergence (KLD).

The Kullback-Leibler divergence measures the relative entropy between two probability distributions. It was defined in information theory (Losee, 1990) and used in many information retrieval and natural language processing tasks, for example, in query expansion following pseudo-relevance feedback (Carpineto et al., 2001).

The KLD score of a subjective lexical unit was calculated according to Eq. 1.

$$KLD(t) = P_R(t) \times \log \frac{P_R(t)}{P_N(t)} \qquad (1)$$

Where: $P_R(t)$ – probability of the subjective lexical unit t occurring in the relevant documents, and calculated as $f_R(t)/R$, where $f_R(t)$ – frequency of occurrence of t in the relevant set, R – number of terms in the relevant set; $P_N(t)$ – probability of the subjective lexical unit t occurring in the non-relevant documents, and calculated as $f_N(t)/N$, where $f_N(t)$ – frequency of occurrence of t in the non-relevant set, N – number of terms in the non-relevant set.

The BLOG track 2006 and 2007 topics were used for calculating KLD scores of the subjective lexicon for the evaluation on 2008 topics, and 2006 topics only were used for calculating KLD scores for the evaluation on 2007 topics. The relevant set consisted of all the documents with the relevance judgments of 2 (negative opinion), 3 (mixed opinion) and 4 (positive opinion); the non-relevant set consisted of all the documents with the judgments of 0 (non-relevant) and 1 (relevant, but not opinionated).

A KLD score was calculated for each lexical unit, not each of its grammatical word forms separately. Thus, in calculating KLD for "fascinate", the frequencies of occurrences of "fascinate", "fascinated", "fascinating" and "fascinates" were added. In total, KLD scores for 6553 lexical units were calculated. Lexical units with negative KLD scores were discarded.

3.5.1.2 Calculating document matching score

Our approach to the weighting of query term occurrences in a document consists of modifying the term frequency (*tf*) calculation in BM25. Instead of counting the actual frequency of a term's occurrence in the document to get *tf*, we calculate a pseudo-frequency (*pf*) value. In this approach, the contribution of each query term instance $c(t_i)$ is not 1 as in *tf*, but can be

zero or greater than 1 depending on a number of factors, such as subjective words in its window and proximity to other query terms. Specifically, if the query term t_i co-occurs with a subjective word within a window of n words either side of it, we add the normalised KLD score of the subjective word to $c(t_i)$; if, however, the query term t_i does not have any subjective words in its window, $c(t_i) = 0$ (Eq. 2). In addition to this, we factor in proximity of the query term to the nearest query term from a different query facet. The idea of pseudo-frequency weights was used earlier in our proximity-based document ranking method proposed in (Vechtomova and Karamuftuoglu, 2008), which proved to be effective in an ad-hoc IR task.

$$c(t_i) = \begin{cases} 1 + FD(t_i) + \sum_{j=1}^{|J|} \frac{KLD(s_j)}{\max KLD} & \text{if } |J| > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Where: $c(t_i)$ – contribution of the i^{th} instance of the query term t to pf, $FD(t_i)$ – facet distance (FD) component, which factors in the distance of t_i to the nearest other query term/phrase from a different facet (Eq. 3), $KLD(s_j)$ – KLD score of the subjective lexical unit s_j, $|J|$ – the number of subjective lexical units occurring in the window of n words around t_i, $maxKLD$ – the maximum KLD score out of all 6553 subjective lexical units.

$$FD(t_i) = \begin{cases} \frac{1}{\sqrt{\min dist(t_i, q)}} & \text{if } q \in D;\ q \in facet_n; t \in facet_m \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Where: min $dist(t_i,q)$ – the distance between t_i and the nearest term/phrase q in the document D, where q belongs to a different query facet than t.

The pseudo frequency pf_t of each query term t is calculated according to Eq. 4.

$$pf_t = \sum_{i=1}^{N} c(t_i) \quad (4)$$

Where: N – number of instances of the query term t in the document.

After pf is calculated for a query term, its Term Weight (TW) in the document is calculated in the same way as in the BM25 formula, with pf used instead of tf (Eq. 5):

$$TW_t = \frac{(k_1 + 1) \times pf_t}{k_1 \times NF + pf_t} \times idf_t \quad (5)$$

Where: k_1 – term frequency normalisation factor, which moderates the contribution of the weight of frequent terms. If k_1=0, pf has no effect on the term weight, while the higher the value of k_1 the more effect pf has on the term weight. NF – document length normalisation factor, and is calculated in the same way as in the BM25 document ranking function, as expressed in Eq. 6.

$$NF = (1 - b) + b \times \frac{DL}{AVDL} \quad (6)$$

Where: b – tuning constant, DL – document length in word counts; $AVDL$ – the average document length in the document collection.

The Document Matching Score is calculated as the sum of weights of all query terms found in the document (Eq. 7).

$$MS = \sum_{t=1}^{|Q|} TW_t \quad (7)$$

Where: $|Q|$ – the number of terms in the query.

The queries used in the opinion re-ranking stage are the same as used in the initial document retrieval stage (baseline), and contain single terms and phrases, identified according to the methods described in Section 3.2, plus the constituent non-stopwords of phrases. For instance, the query for Topic 851 (Title: "March of the Penguins") consists of: "march of the

penguins", "march", "penguins". When calculating weights of single terms, the following rule was followed: if the occurrence of the single term t_i in the document is not a part of the phrase occurrence, its $c(t_i)$ is calculated according to Eq. 2. If, however, it occurs as part of the phrase, then its $c(t_i)$ is calculated according to Eq. 8. The reason for using Eq. 8 in such cases is to avoid double-counting the KLD and FD factors. The same rule was used in the methods described in Sections 3.5.2 and 3.5.3.

$$c(t_i) = \begin{cases} 1 & \text{if } |J| > 0 \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

For example, if a document contains 2 instances of the query phrase "march of the penguins" and 3 separate instances of "penguins", the document matching score will be TW("march of the penguins"[2 instances]) + TW("penguins"[3 instances]) calculated using Eq. 2 + TW("march"[2 instances]) + TW("penguins" [2 instances]) calculated using Eq. 8.

3.5.2 Proximity-based method

In proximity-based term weighting, we also used the idea of calculating a pseudo-frequency (*pf*) value for a query term in a document, which depends on its proximity to each of the instances of subjective lexical units within the window of *n* words (Eq. 9).

$$c(t_i) = \begin{cases} 1 + FD(t_i) + \sum_{j=1}^{|J|} \frac{1}{\sqrt{dist(t_i, s_j)}} & \text{if } |J| > 0 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

Where: $dist(t_i,s_j)$ – distance in number of non-stopwords between the query term t_i and subjective lexical unit s_j; $|J|$ – number of subjective lexical units occurring within the window of n words around t_i.

After $c(t_i)$ is calculated, *pf* and the document matching score are calculated in the same way as described in Section 3.5.1 above.

3.5.3 Combined method

In this method we combined the KLD-based and distance-based term weighting methods as given in Eq. 10.

$$c(t_i) = \begin{cases} 1 + FD(t_i) + \sum_{j=1}^{|J|} \frac{KLD(s_j)}{\max KLD} + \frac{1}{\sqrt{dist(t_i, s_j)}} & \text{if } |J| > 0 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

The document matching score is calculated in the same way as in the previous two methods.

4. Evaluation

The proposed methods were evaluated on 2007 and 2008 Blog track opinion-finding task topics. Two baselines were used for both evaluating our methods and for the initial document retrieval stage of our opinion retrieval algorithm:

1) BM25 implemented in Wumpus, using the limitedQE-Wiki-phrases method described in Section 3.2.3.
2) Baseline 4 provided by the organisers of the Blog track 2008, which facilitates cross-system comparison. The Blog 2008 track organisers released 5 baselines, selected from the baseline runs submitted by track participants, i.e. the runs with all opinion features switched off. Baseline 4 has the highest topic- and opinion-relevance MAP, and therefore serves as a good benchmark to compare our methods to the methods of other Blog 2008 track participants.

In the Blog track, each run can contain up to 1000 retrieved and ranked blog posts per topic. The following naming conventions were used to denote the runs: "KLD" – runs using the KLD-based method (Section 3.5.1), "dist" – runs using the proximity-based method (Section 3.5.2) and "KLD+dist" – runs using the combined method (Section 3.5.3). Suffix "bm25" was appended to the experimental runs based on the BM25 baseline "limitedQE-Wiki-phrases", and suffix "b4" was appended to the experimental runs based on Baseline 4. All experimental runs have facet distance (FD) and facet validation (FV) components.

Table 4 presents the results of opinion-based reranking runs on Blog 2008 topics using the first baseline (BM25). The Kullback-Leibler divergence of subjective words in runs "KLD-FD-FV-subj-bm25" and "KLD+dist-FD-FV-subj-bm25" was calculated using Blog 2006 and 2007 topics. In the table, Δ MAP represents the difference in percentage of the run's MAP from the baseline MAP. Runs marked with * are statistically significant at .05 level; ** at .02 level (paired t-test). All

experimental runs demonstrate statistically significant improvement in most measures. The combined method "KLD+dist-FD-FV-subj-bm25" performs slightly better than either "KLD-FD-FV-subj-bm25", or "dist-bm25" methods in MAP and R-precision, however, the other two runs are slightly better in Precision at 10 documents (P10).

Run	Opinion relevance				Topic relevance			
	MAP_{op}	ΔMAP_{op}	$P10_{op}$	$Rprec_{op}$	MAP_{rel}	ΔMAP_{rel}	$P10_{rel}$	$Rprec_{rel}$
limitedQE-Wiki-phrase (BM25 baseline)	0.3737		0.6180	0.4291	0.4643		0.7280	0.5031
KLD-FD-FV-subj-bm25	0.4017**	7.49%	0.7020**	0.4411	0.4797*	3.32%	0.7820**	0.5094
dist-FD-FV-subj-bm25	0.4008**	7.25%	0.7020**	0.4405	0.4802**	3.42%	0.7820**	0.5083
KLD+dist-FD-FV-subj-bm25	0.4027**	7.76%	0.6920**	0.4464**	0.4828**	3.98%	0.7760**	0.5155

Table 4. Opinion runs with Blog 2008 topics based on "limitedQE-Wiki-phrase" baseline run.

Table 5 shows the results of opinion-based runs on Blog 2007 topics, also using the BM25 baseline. Here, the Kullback-Leibler divergence of subjective words was calculated using only Blog 2006 topics. Again, all runs demonstrate statistically significant improvements over the baseline in most measures. The combined run "KLD+dist-FD-FV-subj-bm25" has the highest results in topic relevance measures, but "KLD-FD-FV-subj-bm25" has the highest MAP_{op} and $P10_{op}$.

Run	Opinion relevance				Topic relevance			
	MAP_{op}	ΔMAP_{op}	$P10_{op}$	$Rprec_{op}$	MAP_{rel}	ΔMAP_{rel}	$P10_{rel}$	$Rprec_{rel}$
limitedQE-Wiki-phrase (BM25 baseline)	0.3759		0.5420	0.4128	0.5249		0.7360	0.5542
KLD-FD-FV-subj-bm25	0.4288**	14.07%	0.6440**	0.4448**	0.5670**	8.02%	0.8460**	0.5799*
dist-FD-FV-subj-bm25	0.4244**	12.90%	0.6100**	0.4477**	0.5659**	7.81%	0.8140**	0.5691
KLD+dist-FD-FV-subj-bm25	0.4262**	13.38%	0.6340**	0.4468**	0.5690**	8.55%	0.8400**	0.5825**

Table 5. Opinion runs with Blog 2007 topics based on "limitedQE-Wiki-phrase" baseline run.

Table 6 shows the opinion-based runs on Blog 2008 topics based on the standard Baseline 4 provided by the track organisers. Again, all experimental runs yield statistically significant improvement in most measures. Here, "KLD+dist-FD-FV-subj-b4" shows slightly higher MAP_{op} and R-precision than the other two experimental runs, whereas they slightly outperform it in $P10_{op}$.

Run	Opinion relevance				Topic relevance			
	MAP_{op}	ΔMAP_{op}	$P10_{op}$	$Rprec_{op}$	MAP_{rel}	ΔMAP_{rel}	$P10_{rel}$	$Rprec_{rel}$
Baseline 4	0.3822		0.6160	0.4284	0.4724		0.7440	0.4993
KLD-FD-FV-subj-b4	0.4213**	10.23%	0.6880**	0.4560**	0.5018**	6.22%	0.7720	0.5256*
dist-FD-FV-subj-b4	0.4226**	10.57%	0.6880**	0.4539**	0.5050**	6.90%	0.7700	0.5275**
KLD+dist-FD-FV-subj-b4	0.4229**	10.65%	0.6840**	0.4601**	0.5047**	6.83%	0.7700	0.5278**

Table 6. Opinion runs with Blog 2008 topics based on "Baseline 4" provided by Blog 2008 organisers.

The proposed three opinion-based re-ranking methods compare favourably to other methods developed by the participants of Blog 2007 and 2008 tracks. In Table 7, we list the 10 best performing opinion runs submitted by participants to Blog 2008 track. The participants' results reported in this table are referenced from (Ounis et al., 2008). Our runs are highlighted in bold. The table reports the results in MAP_{op}, $P10_{op}$ and R-$precision_{op}$. Runs are ranked by MAP_{op}. Most of the participants submitted a corresponding baseline run for their opinion runs, however, some participants did not separate opinion-ranking from topic-ranking features, and did not submit a baseline. ΔMAP_{op} indicates the improvement of each run over its baseline where applicable. Two runs yielded higher MAP_{op} than our methods, however, both of them do not have a corresponding baseline, therefore it is not clear how much of their performance is due to the opinion-finding features or other factors. Our approaches show the best results among the two-stage retrieval methods, which first retrieve documents using topic-based ranking, and then re-rank them using opinion-based methods.

Run	Fields	Baseline	MAP_{op}	ΔMAP_{op}	$P10_{op}$	$Rprec_{op}$
KLEDocOpinT	T	N/A	0.4569	N/A	0.7200	0.4797
top3dt1mRd	T	N/A	0.4335	N/A	0.6780	0.4618
KLD+dist-FD-FV-subj-bm25	**T**	**limitedQE-Wiki-phrase**	**0.4027**	**7.76%**	**0.6920**	**0.4464**
KLD-FD-FV-subj-bm25	**T**	**limitedQE-Wiki-phrase**	**0.4017**	**7.49%**	**0.7020**	**0.4411**
dist-FD-FV-subj-bm25	**T**	**limitedQE-Wiki-phrase**	**0.4008**	**7.25%**	**0.7020**	**0.4405**
DUTIR08Run4	T	DUT08BRun2	0.3902	31.60%	0.6620	0.4257
uams08n1o1sp	T	uams08n1o1	0.3823	0.68%	0.6580	0.4204
uogOPb2ofL	T	uogBLProxCE	0.3709	5.04%	0.6380	0.4049
THUopnTmfRmf	T	THUrelTwpmf	0.3522	6.31%	0.6320	0.4104
UniNEopZ1	TD	UniNEBlog1	0.3418	-4.12%	0.5840	0.3961
prisoa1	T	prisba	0.3344	-0.06%	0.5560	0.3868
DCUCDVPtol	T	DCUCDVPtbl	0.3299	14.75%	0.6360	0.3679
wdqfdt1mRd	TDN	wdoqlnvN	0.3127	13.38%	0.6200	0.3702

Table 7. Our runs (in bold) based on "limitedQE-Wiki-phrase" baseline are compared to the runs using own baselines (or no baseline) submitted by participants to Blog 2008 opinion finding task.

Table 8 compares our methods to the 10 best runs using the standard Baseline 4. Runs are ranked by Δ MAP_{op}. The participants' results reported in this table are referenced from (Ounis et al., 2008). Our methods achieve the highest improvement in MAP_{op}.

Run	Fields	MAP_{op}	ΔMAP_{op}
KLD+dist-FD-FV-subj-b4	**T**	**0.4229**	**10.65%**
dist-FD-FV-subj-b4	**T**	**0.4226**	**10.57%**
KLD-FD-FV-subj-b4	**T**	**0.4213**	**10.23%**
B4PsgOpinAZN	T	0.4189	9.60%
DCUCDVPgoo	TD	0.4155	8.71%
uicop2bl4r	T	0.4067	6.41%
b4dt1mRd	T	0.4023	5.26%
FIUBL4DFR	T	0.4006	4.81%
uogOP4intL	T	0.3964	3.72%
NOpMM47	TD	0.3844	0.58%
UWnb4Op	T	0.3381	-11.54%
KGPBASE4	T	0.2852	-25.38%
uams08b4pr	T	0.1369	-64.18%

Table 8. Our runs (in bold) based on the standard Baseline 4 are compared to the runs using Baseline 4 submitted by participants to Blog 2008 opinion finding task.

We also compare our methods using Blog 2007 topics to the 10 best opinion title-only runs submitted to Blog 2007 opinion finding task (Table 9). The participants' results reported in this table are referenced from (Macdonald et al., 2007). Runs are ranked by MAP_{op}. One run achieved higher MAP_{op} than our methods, but again this is a run that has no corresponding baseline. Our runs were the best among all two-stage retrieval methods.

Run	Baseline	MAP_{op}	ΔMAP_{op}
uic1c	N/A	0.4341	N/A
KLD-FD-FV-subj-bm25	**limitedQE-Wiki-phrase**	**0.4288**	**14.07%**
dist-FD-FV-subj-bm25	**limitedQE-Wiki-phrase**	**0.4244**	**12.90%**
KLD+dist-FD-FV-subj-bm25	**limitedQE-Wiki-phrase**	**0.4262**	**13.38%**
uogBOPFProxW	uogBOPFProx	0.3264	15.87%
DUTRun2	DUTRun1	0.3190	10.38%
FDUTisdOpSVM	FDUNOpRSVMT	0.3179	0.03%
UALR07BlogIU	UALR07Base	0.2911	13.98%
oqsnr2opt	oqsnr1Base	0.2894	14.07%
FIUdPL2	FIUbPL2	0.2728	0.00%
UWopinion3	UWbasePhrase	0.2631	5.83%
NLPRPST	NLPRPTONLY	0.2542	1.44%
EAGLE2	EAGLE1	0.2493	-2.66%

Table 9. Our runs (in bold) based on "limitedQE-Wiki-phrase" baseline are compared to the best runs using own baselines (or no baseline) submitted by participants to Blog 2007 opinion finding task.

5. Analysis and discussion

5.1 Topic-based analysis

Analysis of results by topics shows that the methods improve performance on a large number of topics over the baseline "limited-QE-Wiki-phrase". Among the 2008 topics, "KLD-FD-FV-subj-bm25", "dist-FD-FV-subj-bm25" and "KLD+dist-FD-FV-subj-bm25" improved the opinion relevance Average Precision ($AveP_{op}$) of 35, 36 and 35 topics respectively, and left 1 topic unchanged. Among the 2007 topics, the respective numbers of improved topics are 37, 35 and 38, with 2 topics unchanged. Figure 6 shows the difference in $AveP_{op}$ between the experimental runs and the baseline on 2008 topics.

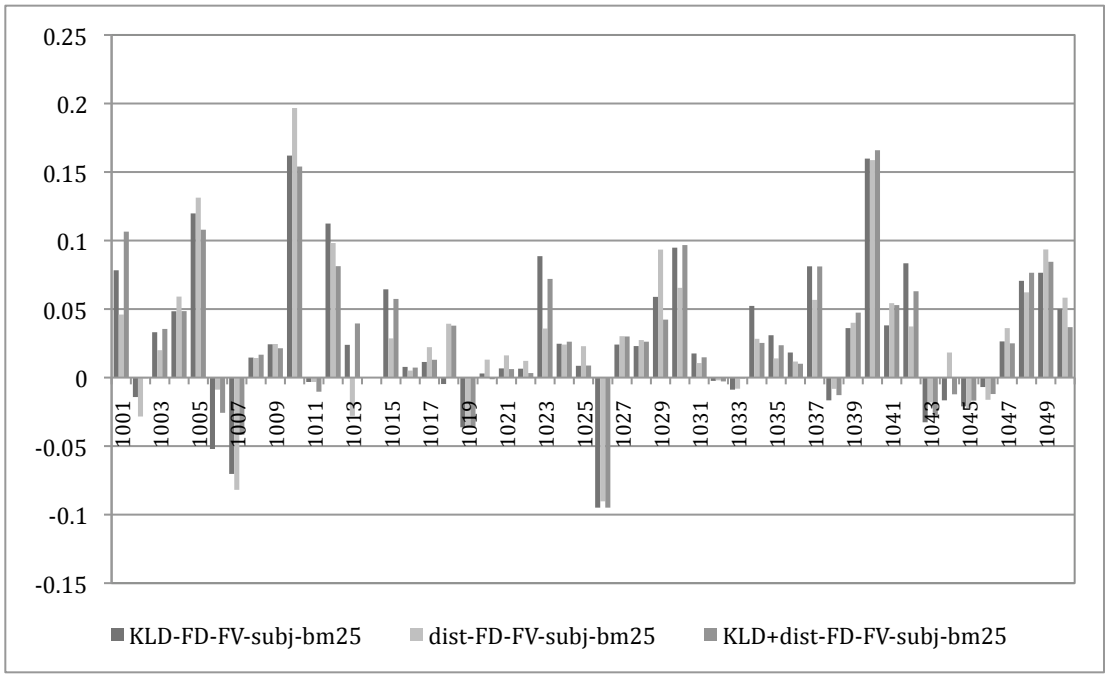

Figure 6. Difference in $AveP_{op}$ between the experimental runs and the baseline "limited-QE-Wiki-phrases" (Blog 2008 topics).

As an example of a document that benefited from the opinion re-ranking methods, consider document "BLOG06-20051208-132-0003415997" judged as 4 (relevant, containing positive opinion) for the Blog 2007 topic 935 "Mozart". The document was retrieved at rank 315 in the baseline "limitedQE-Wiki-phrases" run, but was promoted to ranks 206, 181 and 194 in "KLD-FD-FV-subj-bm25", "dist-FD-FV-subj-bm25" and "KLD+dist-FD-FV-subj-bm25" respectively. An excerpt from the document with the subjective lexicon underlined is shown in Figure 7.

"…The only thing featuring an oboe was a CD of Mozart's Sinfonia Concertante. I loved the piece from the get go. It was so lively yet delicate.

…His intonation was absolutely perfect. Even on that one really high note near the end of the piece. His sound was very beautiful though they were all playing in that sort of bright, Mozart style. What impressed me the most was his expression. Sometimes when you get so used to a specific recording you can't immediately appreciate a different interpretation.

…I really prefer the latter version, but many are now saying the flute version is the more accurate. And some say Mozart didn't write this at all. What did you program notes suggest? How wonderful that you go to hear Klein! I'm quite envious! …"

Figure 7. Excerpt from the document BLOG06-20051208-132-0003415997 (topic 935)

As can be seen from the excerpt, none of the subjective words directly modify the opinion target of the query, "Mozart", but rather words and phrases, such as "a different interpretation", "sound", "intonation", "piece", referring to Mozart's music and its performance. Indeed, people frequently express opinions not directly about the target, but about related concepts. This suggests that further improvements may be obtained by identifying related concepts and possibly performing a more sophisticated discourse analysis, aimed at determining whether the concepts are related contextually.

5.2 The effect of the method components on retrieval performance

We conducted runs with different combinations of components of the proposed methods in order to understand better their effect on the retrieval performance. The components used in each run are given in Table 10.

Run	Facet validation (FV)	Facet distance (FD)	KLD of subjective words(KLD)	Distance to subjective words (dist)	Count only query terms near subj. words (subj)
FV-bm25	✓				
subj-bm25					✓
FV-subj-bm25	✓				✓
FD-FV-subj-bm25	✓	✓			✓
KLD-FD-FV-subj-bm25	✓	✓	✓		✓
dist-FD-FV-subj-bm25	✓	✓		✓	✓
KLD+dist-FD-FV-subj-bm25	✓	✓	✓	✓	✓

Table 10. Components used in the runs

In the run "FV-bm25" the contribution of each query term occurrence to *pf* is 1, regardless of whether it co-occurs with a subjective word in the surrounding window or not. In all the other runs, if a query term occurrence does not have at least 1 subjective word in its window, its contribution to *pf* is 0. Both runs "subj-bm25" and "FV-subj-bm25" use Eq. 11 in calculating the contribution of each query term occurrence to *pf*:

$$c(t_i) = \begin{cases} 1 & \text{if } |J| > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (11)$$

Where: $|J|$ – the number of subjective lexical units occurring in the window of n words around t_i

The run "FD-FV-subj-bm25" uses Eq. 12:

$$c(t_i) = \begin{cases} 1 + FD(t_i) & \text{if } |J| > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (12)$$

The last three runs in Table 10 are the experimental runs presented in the previous section, and are given here for comparison. Table 11 shows the results of the runs on Blog 2008 topics, while Table 12 contains the Blog 2007 results. Runs marked with * are statistically significant at .05 level; ** at .02 level (paired t-test).

Run	Opinion relevance				Topic relevance			
	MAP_{op}	ΔMAP_{op}	$P10_{op}$	$Rprec_{op}$	MAP_{rel}	ΔMAP_{rel}	$P10_{rel}$	$Rprec_{rel}$
limitedQE-Wiki-phrase (BM25 baseline)	0.3737		0.6180	0.4291	0.4643		0.7280	0.5031
FV-bm25	0.3754	0.45%	0.6260	0.4275	0.4647	0.08%	0.7320	0.5008
subj-bm25	0.3776	1.04%	0.6660**	0.4175	0.4519	-2.67%	0.7560	0.4175
FV-subj-bm25	0.3933	5.24%	0.6700**	0.4368	0.4763*	2.58%	0.7620	0.5071
FD-FV-subj-bm25	0.3942	5.49%	0.6760**	0.4372	0.4774	2.82%	0.7660	0.5087
KLD-FD-FV-subj-bm25	0.4017**	7.49%	0.7020**	0.4411	0.4797*	3.32%	0.7820**	0.5094
dist-FD-FV-subj-bm25	0.4008**	7.25%	0.7020**	0.4405	0.4802**	3.42%	0.7820**	0.5083
KLD+dist-FD-FV-subj-bm25	0.4027**	7.76%	0.6920**	0.4464**	0.4828**	3.98%	0.7760**	0.5155

Table 11. Performance of the runs on Blog 2008 topics.

Run	Opinion relevance				Topic relevance			
	MAP_{op}	ΔMAP_{op}	$P10_{op}$	$Rprec_{op}$	MAP_{rel}	ΔMAP_{rel}	$P10_{rel}$	$Rprec_{rel}$
limitedQE-Wiki-phrase (BM25 baseline)	0.3759		0.5420	0.4128	0.5249		0.7360	0.5542
FV-bm25	0.3779	0.45%	0.5460	0.4142	0.5298**	0.93%	0.7400	0.5610
subj-bm25	0.4148**	10.35%	0.6260**	0.4458**	0.5521**	5.18%	0.8200**	0.5708
FV-subj-bm25	0.4180**	11.20%	0.6260**	0.4469**	0.5614**	6.95%	0.8200**	0.5774*
FD-FV-subj-bm25	0.4221**	12.29%	0.6240**	0.4491**	0.5671**	8.04%	0.8180**	0.5824**
KLD-FD-FV-subj-bm25	0.4288**	14.07%	0.6440**	0.4448**	0.5670**	8.02%	0.8460**	0.5799*
dist-FD-FV-subj-bm25	0.4244**	12.90%	0.6100**	0.4477**	0.5659**	7.81%	0.8140**	0.5691
KLD+dist-FD-FV-subj-bm25	0.4262**	13.38%	0.6340**	0.4468**	0.5690**	8.55%	0.8400**	0.5825**

Table 12. Performance of the runs on Blog 2007 topics.

The facet validation (FV) component (run "FV-bm25") and counting only query term instances near subjective words ("subj" component, run "subj-bm25") do not yield substantial improvements each on their own on 2008 topics, however the latter yields statistically significant improvements in all measures except $Rprec_{rel}$ on 2007 topics. The combination of both components (run "FV-subj-bm25") is more effective, yielding 5.24% improvement in MAP_{op} on 2008 topics and 11.2% on 2007 topics over the baseline "limitedQE-Wiki-phrase".

By comparing the runs "KLD-FD-FV-subj-bm25", "dist-FD-FV-subj-bm25" and "KLD+dist-FD-FV-subj-bm25" to the run "FD-FV-subj-bm25", it is possible to understand how much "KLD", "dist" and "KLD+dist" components contribute to performance. Specifically, when "KLD" component is added (run "KLD-FD-FV-subj-bm25"), MAP_{op} improves by 1.9% (statistically significant, $p<0.05$) on 2008 topics, and by 1.59% (not significant) on 2007 topics over "FD-FV-subj-bm25". Adding "dist" component (distance from the query term occurrence to subjective words in its window, run "dist-FD-FV-subj-bm25") improves MAP_{op} by 1.67% and 0.54% (not significant) on 2008 and 2007 topics respectively. Adding both KLD and dist components (run "KLD+dist-FD-FV-subj-bm25") improves MAP_{op} by 2.16% (significant, $p<0.02$) and 0.97% (not significant) on 2008 and 2007 topics respectively.

The facet distance (FD) component (run "FD-FV-subj-bm25") improves performance by 0.23% and 0.98% over "FV-subj-bm25" on 2008 and 2007 topics respectively. In Blog 2008 there are only 10 topics that have more than 1 facet in the query, while in Blog 2007 there are 7 such topics. Table 13 shows results calculated based on only these topics.

Run	Opinion relevance				Topic relevance			
	MAP_{op}	ΔMAP_{op}	$P10_{op}$	$Rprec_{op}$	MAP_{rel}	ΔMAP_{rel}	$P10_{rel}$	$Rprec_{rel}$
Blog 2008 (10 topics):								
FV-subj-bm25	0.3802		0.64	0.4255	0.4402		0.72	0.4909
FD-FV-subj-bm25	0.3845	1.13%	0.67	0.4272	0.4454	1.18%	0.74	0.4999
Blog 2007 (7 topics):								
FV-subj-bm25	0.2859		0.36	0.3075	0.3424		0.53	0.3535
FD-FV-subj-bm25	0.3067	7.28%	0.35	0.3185	0.3709	8.32%	0.52	0.3787

Table 13. Performance of runs with and without the Facet Distance (FD) component on topics with more than one query facet.

5.3 The effect of topic type on performance

Opinion targets in the Blog track topics could be, for instance, people, products, events or abstract concepts. It is interesting to see whether the developed methods are equally effective in finding opinions about different entity types. We manually grouped the topics in Blog 2007 and 2008 into categories based on the type of opinion targets. The categories and their corresponding topic identifiers are given in Table 14.

Category	Topic numbers
Person	903, 904, 908, 920, 922, 924, 935, 940, 941, 947, 949, 1006, 1012, 1017, 1025, 1029, 1034, 1042, 1050
Event	905, 906, 907, 913, 914, 923, 925, 933, 936, 938, 1021
Product	901, 909, 916, 917, 932, 934, 939, 944, 946, 1002, 1005, 1010, 1023, 1024, 1040, 1049
Organisation	910, 912, 915, 919, 926, 930, 937, 948, 950, 1001, 1003, 1004, 1008, 1011, 1013, 1016, 1022, 1027, 1030, 1031, 1033, 1035, 1037, 1038, 1039, 1047
Media/art	911, 921, 928, 1009, 1018, 1032, 1036, 1043, 1045, 1048
Miscellaneous	902, 927, 929, 942, 943, 1007, 1014, 1015, 1019, 1020, 1026, 1028, 1041, 1044, 1046
Geographical location	918, 931, 945

Table 14. Blog 2007 and 2008 topics manually grouped into categories by opinion target type.

MAP_{op} results by opinion target type of the "limitedQE-Wiki-phrase" baseline and the experimental runs are presented in Figure 8. As can be seen from the graph, the category in which the experimental runs achieved the highest improvement over the baseline is "Product", followed by "Geographical location", "Person" and "Organisation". Topics of type "Event" yielded on average the highest MAP in all runs. Topics of types "Media/Art" and "Miscellaneous" appear to have on average no or little benefit from the proposed opinion-finding methods. Topics in the category "Miscellaneous" are mostly abstract concepts, for example, "Oscar fashion" (Topic 927), "talk show hosts" (1044), "tax break for hybrid automobiles" (Topic 1014) and "universal health care" (Topic 1046). Topics, such as 927 and 1044, performed poorly because they use general terms to represent a set of specific entities. In the case of Topic 1044, the searcher is interested in opinions about specific talk show hosts. A relevant document is likely to refer to a talk show and its host by name and may not contain words "talk show" and "host". One possible way to improve the performance of such topics is by expanding the query with the list of names of specific entities (talk shows and their hosts in this example), obtained for example from Wikipedia. As for topics, such as 1014 and 1046, they address rather complex subjects. Many people would express their thoughts and opinions on such topics by using more complex linguistic constructions that are not captured by our methods.

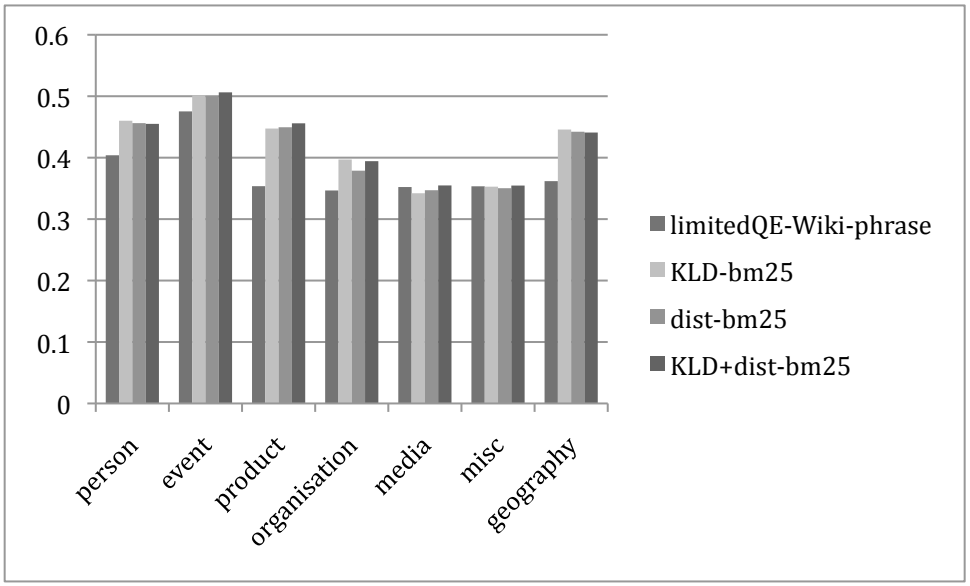

Figure 8. Results in MAP_{op} on Blog 2007 and 2008 topics by opinion target types.

6. Conclusions

In this paper new methods of retrieving documents containing opinions expressed about an entity or entities specified by the user in the query were proposed. The main stages of the proposed methods are as follows:

(1) *Collection pre-processing*. Our experiments demonstrate that this stage has a significant impact on the effectiveness of document retrieval from blogs in terms of both topic- and opinion-relevance. The major performance-improving steps at this stage include the removal of HTML tags, scripts and style definitions, and all lines where the hyperlinks account for 50% or more of the words.

(2) *Query processing*. A method of building faceted queries by utilising Wikipedia was developed. The method consisted of the following steps: identifying concepts in the topic titles by matching them to Wikipedia article titles, grouping concepts into facets, and expanding each facet with new concepts by using Wikipedia article redirects and valid abbreviations.

The evaluation of different query processing levels demonstrates the merit of all three steps in query processing. Expanding queries using only target pages, redirected to from the Wikipedia page titles found in the query ("limitedQE-Wiki-phrases") is better than expanding the query with other pages redirecting to the same targets ("fullQE-Wiki-phrases").

(3) *Document retrieval.* Retrieval of the initial document set using a topic-based ranking method, such as BM25.

(4) *Opinion-based document re-ranking*. Three methods were proposed:

- KLD-based method (KLD), using the Kullback-Leibler divergence scores of the subjective words in the windows around query term occurrences;
- Proximity-based method (dist), using distances between a query term occurrence and each of the co-occurring subjective words;
- A method combining the previous two (KLD+dist).

In addition, all of these methods contain a Facet Distance component, which factors in the distance between query terms/phrases from different facets, and a Facet Validation component, which down-ranks documents that do not contain at least one concept from each facet.

Evaluation demonstrates that the proposed methods are highly effective, and are among the best-performing methods developed by the Blog track 2007 and 2008 participants. Specifically, the proposed methods achieved the highest improvements over the standard baseline run provided by Blog 2008 organisers "Baseline 4" compared to other opinion-finding runs submitted by the participants.

Series of experiments were conducted to determine the effect of the major components (FV, FD, KLD and dist) on performance. The results indicate that all components, in general, have a positive effect on performance. However, the proximity of query terms to subjective words does not always improve the performance when used in conjunction with KL divergence of subjective words. Specifically, "KLD+dist-FD-FV-subj-bm25" yielded lower MAP_{op} and $P10_{op}$ than the run "KLD-FD-FV-subj-bm25" on limitedQE-Wiki-phrase baseline (Blog 2007 topics). "KLD+dist-FD-FV-subj-bm25" and "KLD+dist-FD-FV-subj-b4" on the other hand, yielded higher MAP_{op} and R-$precision_{op}$ on the other two baselines: limitedQE-Wiki-phrase (Blog 2008 topics) and Baseline 4.

An analysis of the methods' performance by topic categories based on the type of entity expressed in the query was performed. It was found that the methods are most effective in finding opinions about events, products, geographical locations and people. They were least effective in finding opinions about entities in the category "media/art", which included TV shows, films and books, and in the category "miscellaneous", which mostly contained abstract concepts.

Among the future extensions of this work, we think that there is a lot of potential in further utilising Wikipedia for finding concepts related to the opinion targets expressed in the query. People commonly express opinions about an entity indirectly, by referring to its related concepts, such as opinions about a composer are expressed by talking about his/her music, or opinions about a company are expressed by discussing its products and services. Our approach to facet-based query structuring and retrieval could be extended further and used to represent more complex queries. For instance, if somebody wants to find opinions about London as a travel destination, they are likely to be interested in opinions about places of interest, museums, hotels, restaurants, etc. A possible structured faceted query could be: Facet 1("London") AND (Facet2("museum", "British museum", "National gallery", "Tate Modern") OR Facet 3("restaurant", "café", "pub")).

References

Amati, G., Ambrosi, E., Bianchi, M., Gaibisso, C. and Gambosi, G. (2008). Automatic Construction of an Opinion-Term Vocabulary for Ad Hoc Retrieval. In Proceedings of ECIR 2008, LNCS vol. 4956, Springer-Verlag Berlin Heidelberg, pp. 89-100.

Baker, C. F., Fillmore, C. J. and Lowe, J. B. (1998). The Berkeley FrameNet project. In Proceedings of COLING-ACL, Montreal, Canada.

Ballmer, Th. and Brennenstuhl, W. (1981). Speech Act Classification. Springer Series in Language and Communication, vol. 8, Springer-Verlag Berlin Heidelberg.

Büttcher, S. and Clarke, C.L.A. (2005). Indexing Time vs. Query Time Trade-offs in Dynamic Information Retrieval Systems. In Proceedings of the 14th ACM Conference on Information and Knowledge Management (CIKM), Bremen, Germany.

Carpineto, C., De Mori, R., Romano, G. and Bigi, B. (2001). An Information-Theoretic Approach to Automatic Query Expansion. ACM Transactions on Information Systems, 19(1), 1–27.

Dave, K., Lawrence, S., and Pennock, D. M. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In Proceedings of the 12th World Wide Web Conference.

Elsas, J.L., Arguello, J., Callan, J., Carbonell, J.G. (2008). Retrieval and Feedback Models for Blog Feed Search. In Proceedings of the International ACM SIGIR Conference on Research & Development on Information Retrieval, pp. 347-354.

Esuli, A. and Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of LREC 2006, pp. 417– 422.

Fellbaum, C. (1998). WordNet An Electronic Lexical Database, MIT Press.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics, pp. 174-181.

He, B., Macdonald, C., Ounis, I., Peng, J. and Santos, R.L.T. (2008a). University of Glasgow at TREC 2008: Experiments in Blog, Enterprise, and Relevance Feedback Tracks with Terrier. Proceedings of the 17th Text Retrieval Conference, Gaithersburg, MD, USA.

He, B., Macdonald, C. and Ounis, I. (2008b). Ranking opinionated blog posts using OpinionFinder. In Proceedings of the International ACM SIGIR Conference on Research & Development on Information Retrieval, pp. 727-728.

Hu, M. and Liu, B. (2004). Mining opinion features in customer reviews. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

Hurst, M. and Nigam, K. (2004). Retrieving topical sentiments from online document collections. In Proceedings of the 11th Conference on Document Recognition and Retrieval.

Jia, L., Yu, C. and Zhang, W. (2008). UIC at TREC 208 Blog Track. In Proceedings of the 17th Text Retrieval Conference, Gaithersburg, MD, USA.

Lee, Y., Na, S., Kim, J., Nam, S., Jung, H. and Lee, J. (2008). KLE at TREC 2008 Blog Track: Blog Post and Feed Retrieval. In Proceedings of the 17th Text Retrieval Conference, Gaithersburg, MD, USA.

Levin, B. (1993). English Verb Classes and Alternations. The University of Chicago Press, Chicago.

Lin, D. (1998). Dependency-Based Evaluation of MINIPAR. In Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation, Granada, Spain.

Losee, R. M. (1990). The science of information: Measurements and applications. Academic Press Prof., Inc., San Diego, CA.

Macdonald, C. and Ounis, I. (2006). The TREC Blogs06 Collection : Creating and Analysing a Blog Test Collection. DCS Technical Report TR-2006-224. Department of Computing Science, University of Glasgow.

Macdonald, C., Ounis, I. and Soboroff, I. (2007). Overview of the TREC-2007 Blog Track. In Proceedings of the 16th Text Retrieval Conference, Gaithersburg, MD, USA.

MacKinnon, I. and Vechtomova, O. (2008). Improving Complex Interactive Question Answering with Wikipedia Anchor Text. In Proceedings of ECIR, LNCS vol. 4956, Springer-Verlag Berlin Heidelberg, pp. 438-445.

Ounis, I., Macdonald, C. and Soboroff, I. (2008). Overview of the TREC-2008 Blog Track. In Proceedings of the 17th Text Retrieval Conference, Gaithersburg, MD, USA.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Riloff, E. and Wiebe, J. (2003). Learning Extraction Patterns for Subjective Expressions. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 105-112.

Spärck Jones, K., Walker, S. and Robertson, S.E. (2000). A probabilistic model of information retrieval: development and comparative experiments. Parts 1 and 2. Information Processing and Management, 36(6), 779-808, 809-840.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), pp. 417-424.

Wilson, T., Wiebe, J. and Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In Proceedings of HLT/EMNLP.

Vechtomova, O. and Karamuftuoglu, M. (2008). Lexical Cohesion and Term Proximity in Document Ranking. Information Processing and Management, 44(4), pp. 1485-1502.

Vechtomova, O. (2007). Using Subjective Adjectives in Opinion Retrieval from Blogs. In Proceedings of the 16th Text Retrieval Conference, November 6-9, 2007, Gaithersburg, MD.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., Patwardhan, S. (2005). OpinionFinder: A system for subjectivity analysis. In Proceedings of HLT/EMNLP 2005 Interactive Demonstrations.

Yang, K. (2008). WIDIT in TREC 2008 Blog Track: Leveraging Multiple Sources of Opinion Evidence. In Proceedings of the 17th Text Retrieval Conference, Gaithersburg, MD, USA.

Yi, J., Nasukawa, T., Bunescu, R., and Niblack, W. (2003). Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In Proceedings of the 3rd IEEE International Conference on Data Mining.