Ad Hoc Retrieval of Documents with Topical Opinion

Jason Skomorowski and Olga Vechtomova

University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada, N2L 3G1
{jcskomor, ovechtom}@uwaterloo.ca

Abstract. With a growing amount of subjective content distributed across the Web, there is a need for a domain-independent information retrieval system that would support ad hoc retrieval of documents expressing opinions on a specific topic of the user's query. In this paper we present a lightweight method for ad hoc retrieval of documents which contain subjective content on the topic of the query. Documents are ranked by the likelihood each document expresses an opinion on a query term, approximated as the likelihood any occurrence of the query term is modified by a subjective adjective. Domain-independent user-based evaluation of the proposed method was conducted, and shows statistically significant gains over the baseline system.

1 Introduction

Users searching for information on the Web may have more complex information needs than simply finding any documents on a certain subject matter. For instance they may want to find documents containing other people's opinions on a certain topic, e.g. product reviews, as opposed to documents with objective content, such as technical specifications. In this work we address the problem of ad hoc retrieval of documents that express opinion on a specific topic. There exist a large number of documents with opinionated content on the Web, however they are scattered across multiple locations, such as individual websites, Usenet groups and web logs ("blogs"). If a person wants to find opinions on a certain subject they have to go to specific websites which might contain such content, for instance, IMDb for film reviews or Amazon for the reviews of books and CDs. Alternatively, they may add words with subjective connotation, such as "review" and "opinion", to their queries. However, it is obvious that only a small fraction of documents expressing opinion on a topic would actually contain words such as "review" or "opinion". There is a clear need for a domain-independent search engine that would support ad hoc retrieval of documents containing opinion about the topic expressed in the user's query. This paper sets to fill this need by proposing a domain-independent method for ad hoc retrieval of documents containing opinion about a query topic.

We propose a lightweight method for ad hoc retrieval of documents which express subjective content about the topic of the query. Documents are ranked by the likelihood each document expresses an opinion on a query term, approximated as the likelihood the query term occurrences in a document are modified by subjective adjectives. For our experiments we use a manually constructed list of subjective adjectives, proposed in [1]. Our method calculates the probability of a noun at a certain distance from an adjective being the target of that adjective. Probabilities at different

G. Amati, C. Carpineto, and G. Romano (Eds.): ECIR 2007, LNCS 4425, pp. 405 – 417, 2007.

distances are precomputed using a parsed training corpus. As part of our approach we have also developed a method of locating a noun modified by an adjective (i.e. resolving an adjective target), which demonstrated high accuracy in our evaluation.

While many elements of language can be used to express subjective content, adjectives are one of the major means of expressing value judgement in English. Our approach of using adjectives as markers of subjective content targeted at the concept expressed in the query relies on the assumption that users frequently want to find opinions about a single entity, such as a product, person, company, travel destination, activity, etc. Such an entity is typically expressed as a noun, a noun phrase or a gerund (a verb with -ing suffix which can act as a noun), and consequently queries of this type consist of either a single query term or a phrase. While it is true that users may be interested in opinions about more complex subjects, such as "The effect of global warming on the environment", opinions on such subjects are likely to be expressed by a greater diversity of more complex language structures (clauses, entire sentences or even paragraphs), and therefore require more sophisticated discourse processing tools. These types of queries are outside the scope of the current work.

In this work we also propose a method of topical opinion ranking by the likelihood a document expresses opinions on the *collocates* of the query terms, i.e. words strongly associated with them in the corpus. The rationale is that an author may express opinion about an entity indirectly, by referring to its related concepts, such as parts or attributes of the car as opposed to the car itself.

The proposed approach is well suited to real-time document retrieval: the computationally expensive task of resolving adjective targets in the training corpus and calculating probabilities of subjective adjectives modifying nouns at various distances is done once at pre-search time, whereas at search time the system only needs to find instances of query terms and subjective adjectives, as well as distances between them.

The rest of the paper is organised as follows: in section 2 we review related work, in section 3 we describe our methodology, including adjective target resolution algorithms and document ranking methods. Section 4 presents evaluation, section 5 discusses the evaluation results, and section 6 concludes the paper and outlines future research directions.

2 Related Work

Although sentiment and subjective language represent a growing research area, work on identifying language that is both subjective and on topic is limited. Hurst and Nigam [2] propose a method of identifying sentences that are relevant to some topic and express opinion on it. First, to determine if a document is relevant to a topic, they use a machine learning approach, trained on hand-labeled documents, 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 polarity (positive or negative). Yi et al. [3] propose to extract 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.

There exists a larger body of research directed towards document classification by sentiment polarity [4, 5, 6, 7]. The focus of these works is on classifying reviews as either positive or negative. A review can be viewed as an example of topical subjectivity with the writer's opinion being a subjective expression on the topic of the item being reviewed. Pang et al. [6] evaluate several machine learning algorithms to classify film reviews as either containing positive or negative opinions. Dave et al. [4] propose and evaluate a number of algorithms for selecting features for document classification by positive and negative sentiment using machine learning approaches. Turney [7] proposes an unsupervised algorithm for classifying reviews as positive or negative. He proposes to identify whether a phrase in a review has a positive or negative connotation by measuring its mutual information with 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 [5] 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. This paper is most closely related to our approach because of its use of adjective proximity.

3 Methodology

In order to determine whether an opinion is given on a topic, we need not only to identify subjectivity in the document, but determine if that subjectivity is being directed at the topic in question. Adjectives have often been defined in terms of their use as a direct noun modifier, and while Baker favours a more general definition for his crosslinguistic study, he agrees that this generalisation holds across "a great many languages" [8]. Not only do adjectives tend to have clear targets, they also are one of the primary means of expressing opinions. While the role played by the adjective can vary widely between languages, value judgement is among the four core semantic types associated with this part of speech [9]. Support for this is found in a study by Bruce and Wiebe which shows the presence of adjectives correlates with subjectivity [10].

The general approach of our work is to rank documents by the likelihood that a document expresses an opinion on a query term, approximating it as the likelihood that the query term occurrences in a document are modified by subjective adjectives. Instead of applying syntactic parsing at search time in order to determine whether a query term instance is the target of a subjective adjective in a document, which is computationally expensive, we instead chose to use a training corpus with marked adjective targets to calculate probabilities that each position outstanding from a subjective adjective contains its target noun. At search time we only have to determine the distance between an instance of the query term and the nearest subjective adjective, and look up the probability that the adjective modifies a noun at this distance. The document score is then calculated as the sum of such probabilities. For this approach we need: a list of subjective adjectives; positional information of index terms in documents; the probability that an adjective modifies a noun at a given distance from it; a corpus where adjectives and their targets are marked for calculating such probabilities.

A list of subjective adjectives can be created manually or automatically, for example, using machine learning techniques. In our work we used a list of 1336 subjective adjectives manually composed by Hatzivassiloglou and McKeown [1]. There also exist many automatic methods of learning subjective language [e.g., 11, 12], which can be used instead. Positional information of index terms in a document is recorded in a typical IR system's index, and therefore is easily obtainable. To calculate the probability that an adjective modifies a noun at a certain distance we need a corpus with marked adjectives and their targets. Such corpus, however, is not available. The method of resolving adjective targets also does not exist. Therefore we developed our own method of resolving adjective targets, which is presented in Section 3.1.

3.1 Resolving Adjective Targets in English

English adjectives are characteristically used either attributively or predicatively [13]. Attributive usage is where a noun is modified directly, typically premodified (e.g., the blue sky). Predicative usage links the adjective to the subject with a copular verb such as "be" (e.g., the sky is blue). Other, less frequent constructions include objective complements of verbs, such as "make" and "prove" (e.g., made the sky blue), resultative secondary predicates [8] (e.g., dyed the sky blue), and degree phrases [14] (e.g., blue as the sky; more blue than the sky).

Since we do not require maximum precision for our application, we will focus our target resolution on only the most frequent usages, attributive and predicative. For identifying resultative secondary predicates we need to have a list of verbs that can be used in such constructs, which is unavailable. Determining the specifics of other usages of adjectives is complicated by the numerous syntactic applications of "as", "than", "make" and other words involved in these constructs.

In order to identify what part of a sentence is being modified by a given adjective, syntactic information is needed. For our approach, we need to know the part of speech (POS) of words and the boundaries of noun phrases, therefore we require a POS tagger and a parser. After evaluating a variety of tools, the SNoW Shallow Parser [15] was found to have a good balance of precision and speed.

3.1.1 Resolving Attributive Use of Adjectives

In the attributive case, a noun phrase to which the adjective refers is the one containing it. In order to determine noun phrase boundaries we use the parser. Manually examining a random sample of 200 subjective adjectives used attributively, we found that the parser fails to find appropriate phrase boundaries in 6.5% of these instances. Most errors involve the parser ending the noun phrase because it has mistagged a noun usage as verb, or erroneously saw an adjective where none exists. A notable limitation of this approach is that it does not account for other noun phrases potentially modified by the adjective via coordinate conjunctions, prepositional phrases, and other constructs. However, it is difficult to identify the correct target in such constructs without the knowledge of their meaning, as demonstrated by the following examples:

- Sudbury is famous for its colourful culture and people. (the people are colourful);
- The artist uses colourful pastels and charcoal. (the charcoal is not colourful);
- A delicious bowl of ice cream. (the ice cream is delicious);
- A ridiculous amount of pasta. (the pasta is not ridiculous).

3.1.2 Resolving Predicative Use of Adjectives

If an adjective occurs outside of a noun phrases, it is likely to be used predicatively. In this case we then read back from the adjective to see if there is a copular verb[1] present before it and, if so, assume the preceding noun phrase to be the subject of that verb and thus predicatively modified by the adjective in question. We employ a variety of measures to improve the accuracy of this approach:

- Only cases where the parser tags the copular verb as actually being used as a verb are considered.
- Clauses delimited from the verb by commas are bypassed when searching for the subject (e.g. The ice-cream, prepared fresh this afternoon, is delicious).
- Situations where there is an intervening noun between the adjective and copular verb are not counted as a predicative construct, because it is most likely that the adjective is used as an objective complement of a different verb (e.g., The ice-cream is made with strawberries and is quite delicious).
- Noun phrases preceded with prepositions are skipped when looking for a potential target as these form a prepositional phrase and are not the subject of the link verb (e.g., The ice-cream in the fridge is old.).

The evaluation of the predicative adjective target resolution algorithm was conducted on the random sample of 200 subjective adjectives used predicatively in the AQUAINT corpus. The target noun phrase was identified correctly in the 86% of cases. The 11% of errors were due to the parser error. One frequent cause of the parser error was that contractions of "not" such as "wasn't" and "didn't" were erroneously tagged as nouns. Only 3% of the errors were caused by our method. While some potential heuristics present themselves, further refinement will be left to later work as additional precision is not necessary to explore search applications and is made irrelevant by parser error.

3.2 Statistics on Adjective Usage

Using the above method and a corpus of text, we can calculate the probability of a noun being the target of an adjective at a certain distance from it. A noun is considered to be the target of an adjective when it is the head of the noun phrase that the adjective modifies as determined by the method described in Section 3.1. We consider the last noun in the noun phrase as the head.

The probability P_i that a noun is the target of (i.e. modified by) an adjective at distance i is calculated according to Eq. 1:

$$P_i = \frac{T_i}{K_i} \tag{1}$$

[1] We used a list of copular verbs from [16].

Where: T_i – the total number of nouns which are targets of any subjective adjective separated by distance i; K_i - total number of nouns separated by distance i from a subjective adjective.

For example, in calculating the probability P_1 that a noun is the target of an adjective which immediately precedes it in text (i.e. noun is located in position 1 relative to the adjective), "deep sea" would count towards T_1 because the adjective "deep" modifies the noun "sea". On the other hand, "exciting sea adventure" would not count towards T_1, because the adjective "exciting" does not modify "sea", but "adventure", which is the head of the noun phrase "sea adventure". Both examples would count towards K_1, because in both of them we have a noun immediately preceded by an adjective.

Table 1. Probabilities of a noun being modified by an adjective at different distances

Distance (i) of noun from adjective	All adjectives			Subjective adjectives		
	Proper nouns	Common nouns	All nouns	Proper nouns	Common nouns	All nouns
-10	0.0026	0.0011	0.0012	0.007	0.0024	0.0026
-9	0.003	0.0016	0.0017	0.0084	0.0033	0.0036
-8	0.0037	0.0021	0.0022	0.0098	0.0048	0.0051
-7	0.0052	0.0031	0.0032	0.0141	0.0068	0.0072
-6	0.0073	0.0045	0.0047	0.0194	0.01	0.0105
-5	0.0112	0.0065	0.0069	0.031	0.0147	0.0156
-4	0.0206	0.0105	0.0112	0.061	0.025	0.027
-3	0.0414	0.0218	0.0232	0.1265	0.0545	0.0585
-2	0.0568	0.0294	0.0313	0.1657	0.0712	0.0765
-1	0.0077	0.0029	0.0033	0.0068	0.0014	0.0017
1	0.331	0.6689	0.6451	0.1971	0.5886	0.5666
2	0.1775	0.1741	0.1743	0.1283	0.1517	0.1504
3	0.1761	0.0489	0.0579	0.1133	0.04	0.0441
4	0.0911	0.0143	0.0197	0.0441	0.0123	0.0141
5	0.0326	0.0041	0.0061	0.017	0.0034	0.0042
6	0.0109	0.0014	0.0021	0.0073	0.0011	0.0014
7	0.0041	0.0005	0.0008	0.0028	0.0004	0.0005
8	0.0022	0.0002	0.0004	0.0021	0.0002	0.0003
9	0.0012	0.0001	0.0002	0.0013	0.0001	0.0001
10	0.0004	0.0001	0.0001	0.0002	0	0

Table 1 contains the probabilities of nouns which, at some position relative to an adjective at position 0, are the target of that adjective. We only calculated probabilities for positions of +/-10 words away from an adjective, based on the average sentence size of 21 words. The probabilities were calculated from the AQUAINT corpus.

As can be seen from Table 1, the position immediately following a subjective adjective (position 1) has the highest probability (0.5666) of containing the target of the adjective (see the last column of Table 1). Position with the next highest probability of containing the target is one word away following the adjective (position 2), which is due to the cases where the target is the head of a longer noun phrase with an intervening modifier. Position -2 has the next highest probability of containing the target noun of a subjective adjective, which represents predicative use of adjectives.

Out of all adjectives, 77% are used attributively and 9% predicatively. When restricted to subjective adjectives, the count becomes 65% attributive and 20% predicative. One explanation for the larger proportion of subjective adjectives used predicatively compared to all adjectives may be that subjectivity is more often directed at proper nouns. Proper nouns do not usually take prenominal adjectives [17], so this attributive usage would need to be written predicatively instead (e.g., one is more likely to say "tall person" or "Jane is tall", but less likely "tall Jane").

3.3 Document Ranking

The goal of our method is to rank documents by the likelihood that they express opinions on the query concept. Our method, therefore, attempts to rank documents by topical subjectivity, i.e. expression of opinion about the query topic. Document ranking is performed by locating all instances of subjective adjectives[2] in the document and computing the aggregate probability that they refer to occurrences of the query term based on the precomputed probabilities described in the previous section.

In more detail a document score is calculated as follows: first the user's query term (or phrase) is used to find a set of top N ranked documents using a best-match IR system. In each document all instances of the query term and subjective adjectives are identified. For each occurrence of the query term, we determine if there are any subjective adjectives within 10 words either side, and note the distance separating them. The probability of a subjective adjective referring to a query term instance occurring i words away is referenced from precomputed statistics[3] (Table 1). We use the sum of these probabilities as the probability that the document contains an opinion on the query topic. The sum of probabilities is calculated according the inclusion-exclusion formula [18] for n non-mutually exclusive events (Eq. 2):

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i} P(A_i) - \sum_{i<j} P(A_i A_j) + \sum_{i<j<k} P(A_i A_j A_k) - ... + (-1)^{n+1} P(A_1 ... A_n) \qquad (2)$$

Where, A_i – co-occurrence of a query term with a subjective adjective at distance i in the document; $P(A_i)$ – precomputed probability (from Table 1) that at distance i a subjective adjective modifies a noun.

The instance of the inclusion-exclusion formula for three events (i.e. three *query term – subjective adjective* co-occurrence pairs) is presented in Eq. 3:

$$P(A_i \cup A_j \cup A_k) = P(A_i) + P(A_j) + P(A_k) - P(A_i A_j) - P(A_i A_k) - P(A_j A_k) + P(A_i A_j A_k) \qquad (3)$$

3.4 Collocates of Query Terms as Opinion Targets

A document can express opinion not directly about the concept represented by the query term, but about related concepts, which can be more general or specific. For example an author may talk subjectively about a film by expressing opinions on the actors' performance or a particular scene or work of the director in general. Another example would be

[2] We used a list of manually tagged subjective adjectives from [1].

[3] In the evaluation we used proper noun statistics as most of the user queries were proper nouns.

someone giving a review of an automobile model by talking about its specific features or components, such as fuel efficiency, comfort, engine or accumulator.

In this work we propose a method of using collocates, words significantly co-occurring in the contexts of query terms in the corpus, as representatives of concepts related to the query topic. Specifically, our approach consists of first finding collocates of a query term, and then calculating a document score which is an aggregate probability that subjective adjectives modify the original query term instances plus instances of their collocates. The next section describes the method used for collocate selection.

3.4.1 Collocate Selection Method

A large number of statistical methods have been used to find and rank collocates, such as Mutual Information [19], Z-score [20], Log-Likelihood ratio and chi-square test [21]. We can view the problem of finding related terms for opinion scoring as similar to query expansion. The difference is that we do not explicitly add additional terms to the query, but use their probabilities of being the target of a subjective adjective as additional evidence that the document expresses opinion on the query topic.

It is outside of the scope of the present work to evaluate different term association measures, therefore we chose to use one term association measure, Z-score, which showed good performance in query expansion experiments [20]. Systematic comparison of different term selection measures is left for future work. Z-score is a statistic for hypothesis testing, i.e. for assessing whether a certain event is due to chance or not. When used for collocation selection, Z-score tests whether the co-occurrence of two words is due to other factors than chance. It is similar to a t-score measure as proposed by Church et al. [19].

We used the method for extracting collocates and calculating Z-score as proposed in [20]. The procedure and parameters we used for selecting collocates are as follows: in the 50 top ranked documents retrieved in response to the user's query term, all terms surrounding instances of the query term within the windows of 20 words (10 words either side of the query term instance) are extracted. In cases where windows surrounding query term instances overlap, terms are extracted only once. All extracted terms are then ranked according to the modified Z-score in Eq. 4 [20], and up to 12 top-ranked terms are selected for the use in our method. All collocates with Z-score less than the significance threshold of 1.6 were rejected.

$$Z = \frac{f_r(x,y) - \dfrac{f_c(y) f_r(x) v_x(R)}{N}}{\sqrt{\dfrac{f_c(y) f_r(x) v_x(R)}{N}}} \tag{4}$$

Where: R – the set of top retrieved documents; $f_r(x,y)$ – joint frequency of x and y in R; $f_c(y)$ – frequency of y in the corpus; $f_r(x)$ – frequency of x in R, $v_x(R)$ – average window size around x in the relevant documents; N – corpus size.

More information about the modified Z-score and its derivation can be found in [20]. The values chosen for the parameters in our study (the window size and the number of Z-ranked collocates selected) are those that showed best results in the query expansion experiments by [20]. It is left for future work to systematically evaluate which parameters perform best in our task.

Table 2 shows a list of collocates selected for the sample of queries submitted by users in our evaluation experiment, which will be described in the next section.

Table 2. Collocates selected for a sample of queries submitted by users in the evaluation experiment

Bill Gates	wealth, dynamite, Microsoft, rich, interview, Napoleon, dollars, he, man, person, short, say
Egypt	ancient, guardian, pyramids, tour, arab, egyptian, travel, Nile, Cairo, modern, country, history
J.K. Rowling	Bloomsbury, Potter, Harry, author, interview, book, books, site
JDeveloper	Oracle, 10g, soa, oc4j, Webgalileo, Oracle9i, ide, bpel, clover, jsf, adf, java
iPod	nano, Apple, iTunes, grayscale, 30gb, generation, mini, dock, gb, shuffle, applecare, playback

3.4.2 Document Ranking Using Collocates

After the set of collocates of the query term is selected, a score is calculated for each of the top N documents retrieved in response to the user's query as follows: in each document all instances of the query term, collocates and subjective adjectives are identified. For each occurrence of the query term and collocates, determine if there are any subjective adjectives within 10 words and note the distance separating them. For each subjective adjective get the probability from precomputed statistics (section 3.2) that it refers to a query term or a collocate instance occurring i words away. Aggregate probabilities are calculated according to Eq. 2 (section 3.3).

4 Evaluation

We conducted a user-based evaluation of the proposed approach. Altogether 33 users, solicited from the University of Waterloo graduate student mailing list, voluntarily participated in the evaluation. The form requesting users to submit their queries contained the following instructions:

"Please enter a word or phrase identifying some person/item/entity, about which you are curious to see opinions. This should complete the sentence: "I'd like to know what the Web thinks of ____________".

The form also contained a text field where users were asked to enter a more detailed description of their information need for future analysis.

We used Google to retrieve the initial set of documents in response to the users' queries. The retrieved documents consist of all results obtainable via the Google API up to one thousand. Because of the limit on the number of documents that Google can return per day via its API, it was not possible to simulate the search process in real time. Users were therefore asked to come back in a few days after query submission in order to do the relevance judgements. Each user submitted one query, and in total for 33 queries 1192 documents were judged.

We evaluated two methods of ranking documents by topical opinion: "Opinion" method using only original query terms (section 3.3); "Collocation opinion" method using original query terms plus their collocates (section 3.4.2).

The baseline against which the above two methods are evaluated is the original Google ranking. The reason for selecting Google as the baseline is that it is one of the most widely used state-of-the-art Web search engines, which users may use to satisfy their opinion information needs. Also, to our knowledge there is no publicly available Web search engine which specifically retrieves documents containing opinions on the subject of the user's query.

For each topic, up to the top 1000 documents retrieved by Google are re-ranked using "Opinion" and "Collocation Opinion" methods. Top 15 ranked documents from each of the three ranked document sets, the "Opinion", "Collocation Opinion" and Google, are extracted, and presented in the random order to the user. By randomizing the order in which documents in the three sets are presented to the user, we ensure that the user is unable to infer which method was used to retrieve each document. It also removes the possibility of user bias due to ranking order. The decision of including only the top 15 documents from each of the three retrieved sets in the results list was made so that the relevance judgement task was not too time-consuming for users. Users were asked to judge the full text of each document in the list as one of the following:

1. Containing an opinion about the query topic ("query relevance");
2. Containing an opinion about something closely related to the query topic ("relevance to a related topic");
3. Containing no opinion about the query or related topics.

5 Results

The performance of two methods "Opinion" and "Collocation Opinion" was evaluated by means of Precision at 10 retrieved documents (P@10) and Precision at 15 retrieved documents (P@15) using "query relevance" and "relevance to a related topic" judgements. Results are presented in Table 3.

Table 3. Evaluation results (* indicates that a run has a statistically significant difference from the baseline, paired t-test, P<0.05)

Method	Query relevance		Relevance to a related topic	
	P@10	P@15	P@10	P@15
Google (baseline)	0.3758	0.3717	0.5424	0.5455
Opinion	0.5182*	0.4990*	0.6636*	0.6626*
Collocation opinion	0.4727*	0.4747*	0.6363*	0.6404*

As can be seen from Table 3 all runs significantly (paired t-test, P<0.05) improved performance of topical opinion retrieval over the baseline. The use of only query terms in estimating the likelihood of the document expressing opinion on the query topic performs better than the use of collocates. Using query relevance judgements, out of 33 queries, "opinion" ranking method improves P@15 of 24, deteriorates 6, and does not affect 3 queries, while "collocation opinion" method improves P@15 of

21 queries, deteriorates 6, and does not affect 6 (Fig. 1). In P@10 "opinion" method improves performance in 21 cases, deteriorates 6, and does not affect 6, while "collocation opinion" improves the performance of 20 queries, deteriorates 10, and does not affect 3. Using related topic judgements, "opinion" improves P@15 of 24, deteriorates 6, and does not affect 3 queries, while "collocation opinion" method improves P@15 of 23 queries, deteriorates 7, and does not affect 3 queries. In P@10 "opinion" method improves performance in 22 queries, deteriorates 6, and does not affect 5. "Collocation opinion" also improves P@10 of 22 queries, but deteriorates 11.

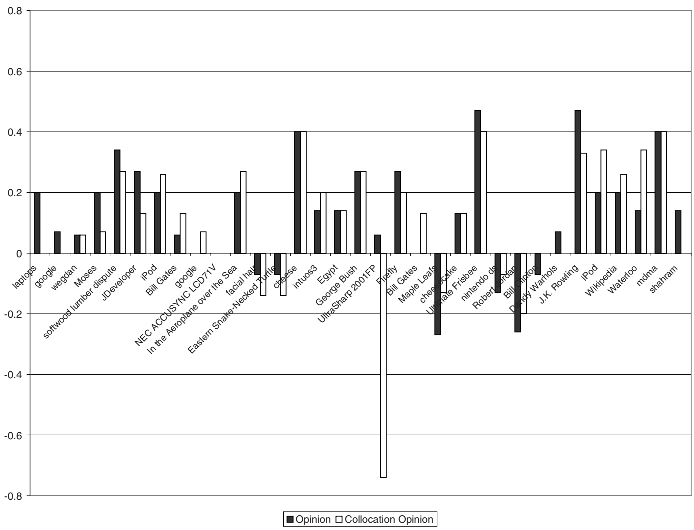

Fig. 1. Precision at 15 (P@15) results for Opinion and Collocation Opinion ("Query Relevance" judgements) relative to the baseline

Our analysis also shows that "collocation opinion" method has better P@10 than "opinion" in 8 queries in the "query relevance" judgements, however it has better P@10 in 10 queries in the "related topic relevance" judgements. This suggests that the "collocation opinion" method may be more suitable for situations where the user's information need is broad, and she may be interested in documents expressing opinions on subjects related to the query topic.

As can be seen from Fig. 1, the topics that benefit most from the “Collocation Opinion” method are those where the author is likely to express opinion about related concepts, such as features or characteristics of the topic entity. For instance, the topic “iPod” performed better with the use of collocates. Some of the collocates selected for this topic represent features of iPod, namely “playback” and “shuffle”, and it is likely that authors express their opinion about iPod through such features. Examples of other topics that performed better with “Collocation Opinion” are “Google”, “Wikipedia”, “Intuos3” and “Bill Gates”.

6 Conclusions and Future Work

In this paper we proposed a computationally lightweight algorithm for topical opinion retrieval. As an element of our technique, we developed a method for adjective target resolution in English, which demonstrated high accuracy. We conducted a thorough user-based evaluation of the developed method in an unrestricted domain using Web as the corpus. Comparison of the developed method to a state-of-the-art search engine (Google) shows statistically significant gains, demonstrating that the system is useful in resolving genuine topical opinion needs of real users. Since at present no domain-independent topical opinion search engine exists, our experiment demonstrates potential uses of such a system, and the types of queries that people may ask.

In addition to the subjectivity of adjectives, the system could incorporate additional metainformation on each adjective, including polarity (positive and negative) and intensity. This would enable more expressive queries to be formulated, limiting which subset of adjectives is applied. For example, a company might be most interested in the superlatively negative comments about its brand, or a consumer might prefer a balance of both positive and negative opinions to find more thorough product evaluations. A metric for opinion quality is one direction for this line of research and could incorporate other indicators of a substantiated rather than casual opinion. We plan to evaluate the above methods and their further extensions by means of user-based evaluations and test collections such as the one created in the Blog track of TREC.

References

1. V. Hatzivassiloglou and K. R. McKeown, Predicting the semantic orientation of adjectives. In Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics, pp. 174-181, 1997.
2. M. Hurst and K. Nigam, Retrieving topical sentiments from online document collections. In Proceedings of the 11th conference on document recognition and retrieval, 2004.
3. J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In Proceedings of the 3rd IEEE International Conference on Data Mining, 2003.
4. K. Dave, S. Lawrence, and D. M. Pennock, Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In Proceedings of the 12th World Wide Web Conference, 2003.
5. M. Hu and B. Liu, Mining opinion features in customer reviews. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2004.
6. B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the 2002 conference on empirical methods in natural language processing, 2002.
7. P. D. Turney, 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, 2002.
8. M. C. Baker, Lexical categories: verbs, nouns and adjectives. Cambridge University Press, 2003.
9. R. M. W. Dixon and A. Y. Aikhenvald, Adjective classes. A crosslinguistic typology. Oxford University Press, 2004.

10. R. F. Bruce and J. M. Wiebe, Recognizing subjectivity: a case study in manual tagging. Natural Language Engineering, vol. 5, no. 2, pp. 187-205, 1999.
11. J. Wiebe, Learning subjective adjectives from corpora. In Proceedings of the 17th National Conference on Artificial Intelligence, pp. 735-740, 2000.
12. P. D. Turney and M. L. Littman, Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Tech. Rep. EGB-1094, National Research Council Canada, 2002.
13. S. Greenbaum, The Oxford English grammar. Oxford University Press, 1996.
14. P. Rijkhoek, On Degree Phrases and Result Clauses. PhD thesis, University of Groningen, Groningen, 1998.
15. X. Li and D. Roth, Exploring evidence for shallow parsing. In Proceedings of the Annual Conference on Computational Natural Language Learning, 2001.
16. J. Sinclair (Ed.) Collins Cobuild English Grammar. Harper Collins, 1990.
17. Z. Vendler, Adjectives and nominalizations, p. 86. Mouton & Co. N.V., The Hague, 1968.
18. J. Pitman, Probability, p. 559. Springer-Verlag, New York, 1993.
19. K. Church, W. Gale, P. Hanks, and D. Hindle, Lexical substitutability. In: Atkins B.T.S. and Zampoli A. (eds.) Computational Approaches to the Lexicon. Oxford University Press, 1994. pp. 153-177.
20. O. Vechtomova, S. E. Robertson, and S. Jones, Query expansion with long-span collocates. Information Retrieval, vol. 6, pp. 251-273, 2003.
21. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, Massachusetts, 1999.