Topical Opinion Retrieval

by

Jason Skomorowski

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Mathematics

in

Computer Science

Waterloo, Ontario, Canada, 2006

Author’s Declaration for Electronic Submission of a Thesis

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

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. While the research area of opinion detection and sentiment analysis has received much attention in the recent years, little research has been done on identifying subjective content targeted at a specific topic, i.e. expressing topical opinion. This thesis presents a novel 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 methods was conducted, and shows statistically significant gains over Google ranking as the baseline.

Acknowledgements

The patience and support of my supervisor, family and friends goes beyond what can be expressed here and has instilled lasting gratitude.

I also wish to thank Mark Sammons at UIUC for assistance with the SNoW parser and provision of handy scripts.

Contents

List of Tables

List of Figures

Chapter 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 thesis sets to fill this need by proposing a domain-independent method for ad hoc retrieval of documents containing opinion about a query topic.

Recent years have seen a growing interest in the detection of subjective content in text, which has led to a number of approaches (Dave et al, 2003; Hu and Liu 2004; Pang et al., 2002; Turney 2002), which will be discussed in more detail in the next section. However, the majority of these works are based on the assumption that if a document is on topic and contains subjective language, it expresses opinion on this topic. To our knowledge only two works (Hurst and Nigam, 2004 and Yi et al., 2003) tried to explicitly link subjectivity with topicality, although their methods were applied to document classification, rather than ad hoc retrieval.

In this thesis a lightweight method for ad hoc retrieval of documents which express subjective content *about* the topic of the query is proposed. Our approach is to rank documents by the likelihood each document expresses an opinion on a query term, approximating it 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 (Hatzivassiloglou and McKeown, 1997), however as will be discussed in section 2.3 there exist many automatic methods of learning subjective language, which can be used instead. Our method calculates the probability of a noun at a certain distance from an adjective being the target of that adjective. Probabilities at different 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 language expressions 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 (e.g., hiking), 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 thesis is organised as follows: Chapter 2 contains the review of related work; the developed methods are described in Chapter 3, including adjective target resolution algorithms and document ranking methods; Chapter 4 presents evaluation; Chapter 5 discusses the evaluation results; Chapter 6 concludes the thesis and outlines future research directions.

Chapter 2

Related Work

2.1 Information Retrieval

The research area of information retrieval or IR is broad and loosely defined despite having what would appear to be a straightforward name. Information is not only text, it is available in myriad forms such as human DNA or a television programme. Retrieval can be more than just matching on simple criteria, it may take into account knowledge of past queries or produce a summary by aggregating many sources. The area also encompasses the organisation and storage techniques that make the information available for such analysis. In approaching these varied challenges, information retrieval draws from diverse fields including computer science, information science, library science, artificial intelligence, cognitive psychology, statistics and linguistics.

One of the information retrieval tasks is ad hoc document retrieval. Concerned with satisfying an information need from a collection of documents, with only a query used to predict which are relevant, ad hoc retrieval refers to the task typically performed by users searching for information on the Web by means of search engines like Google or Alta

Vista. The most salient aspect of any ad hoc retrieval technique is the formula used to retrieve and rank documents.

One of the simplest effective models to rank documents in IR is tf.idf or term frequency–inverse document frequency. With this measure, a document’s relevance is predicted by how frequently the query term appears in the document, weighted by how rare the query term is in the collection i.e. the inverse document frequency. The tf.idf weight of a term is calculated according to Eq. 1:

$$w_i = tf_i \times \log_2 \frac{N}{n_i} \qquad (1)$$

Where: tf_i – frequency of term i in a document; N – number of documents in the collection; n_i – number of documents in the collection containing the term i.

One other well-know model of IR which showed consistently good performance in large-scale evaluations conducted within the framework of the Text Retrieval Conference (TREC) (Voorhees and Buckland 2004) is the Robertson/Sparck Jones probabilistic model (Robertson et al., 1995; Sparck Jones et al. 2000). This model ranks documents by the probability that a document is relevant to the query. It applies Bayes’ Rule to estimate the probability that each query term occurs in the known relevant and non-relevant documents, allowing known relevant documents to inform the process. Without relevance information, the probabilistic model is equivalent to tf.idf. The term weighting function based on this model, BM25 (Sparck Jones et al. 2000), includes additional weighting factors based on document length and term frequency within the document and the query (Eqs. 2, 3).

$$w_i = \frac{tf_i(k_1+1)}{K+tf_i}\log\frac{(r_i+0.5)(N-n_i-R+r_i+0.5)}{(R-r_i+0.5)(n_i-r_i+0.5)} \qquad (2)$$

Where, N is the total number of documents in the corpus; n_i – number of documents containing the term i; R – number of known relevant documents in the corpus; r_i – number of relevant documents containing the term i; tf_i – term frequency of the term i in the document; k_1 - constant moderating the effect of term frequency on the term weight (k_1=0 reduces the term weight to the term-presence weight only, whereas large k_1 values make the term weight nearly linear to tf[1]); K – see Eq. 3 below.

$$K = k_1 \times ((1-b) + b\frac{DL}{AVDL}) \qquad (3)$$

Where, b – constant moderating the effect of document length normalization (smaller values reduce the normalization effect); DL – length of the document in stems (excluding stopwords); $AVDL$ – average length of documents in the collection.

Our ad hoc retrieval approach is also probabilistic in nature. However, rather than estimating the probability of a document's relevance to the query, this work focuses on estimating the probability a document contains an opinion on the query.

[1] Values of 1.2-2 for k1 and 0.75 for b have been found optimal in TREC (Text Retrieval Conference) evaluations (Sparck Jones et al. 2000)

2.2 Topical Subjectivity

Although sentiment and subjective language represent a growing research area, work on identifying language that is both subjective and on topic is very limited. Hurst and Nigam (2004) 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 (Winnow classifier), 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). They evaluated their classification method on a set of 982 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) 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 the mixture language model and 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%.

There exists a larger body of research 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. 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. (2002) evaluate several machine learning algorithms to classify film reviews as either containing positive or negative opinions. Dave et al. (2003) propose and evaluate a number of algorithms for selecting features for document classification by positive and negative sentiment using machine learning approaches. Turney (2002) 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 (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. This paper is most closely related to our approach because of its use of adjective proximity.

2.3 Learning Subjective Language

To work with subjective adjectives, one needs to determine which adjectives are subjective. The subjectivity and polarity (semantic orientation) of language has been investigated at some length. Hatzivassiloglou and McKeown (1997) distinguish between positive and negative adjectives through the hypothesis that adjectives joined by the conjunction "and" tend to be similar, and dissimilar if joined by "but". Wiebe (2000) expands a seed set of manually tagged adjectives by generating a list of correlated words for each seed term using a mutual information metric.

Turney and Littman (2002) define sets of positive and negative adjectives. They determine the orientation of a term based on its association with these sets, exploring both Pointwise Mutual Information and latent semantic analysis for this purpose.

Whitelaw et al. (2005) propose a method of heuristically extracting adjectival appraisal groups, consisting of an appraising adjective (e.g. “beautiful”) and optional modifiers (e.g. “very”). They developed several taxonomies of appraisal attributes by semi-automatically classifying 1329 adjectives and modifiers.

Research on subjective language extends beyond adjectives. Bethard et al. (2004) posit that for question answering applications it may be more appropriate to work with propositional phrases and experiment with using support vector machines to extract them.

Rather than a simple vocabulary vector, they train on specific features derived from parse trees.

Wiebe et al. (2004) derive a variety of subjectivity cues from corpora and demonstrate their effectiveness on classification tasks. They determine a relationship between low-frequency terms and subjectivity and find that their method for extracting subjective n-grams is enhanced by examining those that occur with unique terms.

Esuli and Sebastiani (2005) present a semi-supervised method for identifying the semantic orientation of words using their gloss definitions from online dictionaries. A manually composed seed set of words with positive and negative connotation is provided as input, which is expanded with the words' synonyms from an online dictionary. A text classifier is trained to predict the polarity of words on the basis of their glosses.

Kim and Hovy (2006) proposed a method of automatically expanding a set of seed words, (adjectives and verbs) which were manually tagged as having positive, negative and neutral polarity, with their synonyms identified using WordNet (Miller, 1990). They acknowledged that synonyms of a word may not all have the same polarity, and proposed a method to calculate the closeness of a synonym to each polarity category in order to determine the most probable one. The method was evaluated on a corpus of German emails, and achieved 77% accuracy on verbs, and 69% on adjectives.

Chapter 3

Methodology

3.1 Introduction

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 (2003) favours a more general definition for his crosslinguistic study, he agrees that this generalisation holds across "a great many languages". 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 (Dixon and Aikhenvald, 2004). Support for this is found in a study by Bruce and Wiebe (1999) which shows the presence of adjectives correlates with subjectivity.

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 the following data:

- 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 using machine learning techniques as described in section 2.2 above. In our work we used a list of 1336 subjective adjectives manually composed by Hatzivassiloglou and McKeown (1997). 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 have developed our own method of resolving adjective targets, which is presented in section 3.2.

3.2 Resolving Adjective Targets in English

English adjectives are characteristically used either attributively or predicatively (Greenbaum, 1996). 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 (Baker, 2003), e.g., "dyed the sky blue", and degree phrases (Rijkhoek, 1998), 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 from the University of Illinois (Li and Roth, 2001; Munoz et al., 1999) was found to have a good balance of precision and speed.

3.2.1 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).

3.2.2 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[2] present before it and, if so, estimate 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.

3.2.3 Finding the Head of the Noun Phrase

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 3.2. Since the SNoW parser separates postmodifying clauses, we can ~~We~~ consider the last noun in the noun phrase as the head.

3.3 Statistics on Adjective Usage

The probability that an adjective modifies a noun at a distance *d* is calculated according to Eq. 4:

$$P = \frac{T_d}{K} \qquad (4)$$

[2] We used a list of copular verbs from (Sinclair, 1990).

Where: T_d – the total number of nouns which are targets of any subjective adjective separated by a distance d; K - total number of nouns separated by a distance d from a subjective adjective.

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.

Distance (*d*) 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. Probabilities of a noun being modified by an adjective at different distances.

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 (Vendler, 1968), 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").

3.4 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[3] 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 *p(A$_d$)* of a subjective adjective referring to a query term instance occurring *d* words away is referenced from precomputed statistics[4] (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 to the inclusion-exclusion formula for *n* non-mutually exclusive events (Eq. 5):

$$P(\bigcup_{i=1}^{n} A_i) = \sum_i P(A_i) - \sum_{i<j} P(A_i A_j) + \sum_{i<j<k} P(A_i A_j A_k) - ... + (-1)^{n+1} P(A_1 ... A_n) \quad (5)$$

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. 6:

$$P(A_i \cup A_j \cup A_k) = P(A_i) + P(A_j) + P(A_k) - P(A_i A_j) - P(A_i A_k) - P(A_j A_k) + P(A_i A_j A_k) \quad (6)$$

3.5 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

[3] We used a list of manually tagged subjective adjectives from (Hatzivassiloglou and McKeown, 1997).

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

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

The sources of words representing related concepts can be either human-engineered knowledge bases, such as thesauri, lexical resources (e.g. WordNet) and ontologies, or words extracted using corpus-based statistical and NLP (Natural Language Processing) methods. While methods relying on human-engineered resources may have an advantage of yielding fewer noise terms for some queries, their main disadvantage is that they are either domain-specific, or do not have sufficient lexical coverage. For example, WordNet does not contain proper nouns. Knowledge-based methods also require substantial human effort to construct and keep up to date. 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.

3.5.1 Collocate selection method

A large number of statistical methods have been used to find and rank collocates, such as Pointwise Mutual Information (Church et al., 1994), Z-score (Vechtomova et al., 2003), Log-Likelihood ratio and chi-square test (Manning and Schütze, 1999). In IR collocations have been used in query expansion (Harper and Rijsbergen, 1978; Smeaton and Rijsbergen, 1983; Rijsbergen, 1977; Vechtomova et al., 2003). 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.

In addition to collocation measures, research in query expansion has generated a large number of statistical, NLP and combination methods for selecting query expansion terms and phrases, e.g., (Xu and Croft, 1996). In our approach we opted to use statistical techniques rather than relying on NLP tools, such as POS tagging and parsing, due to computational expensiveness of the latter.

It is outside of the scope of the present work to evaluate different term association and query expansion measures for our task, therefore we chose to evaluate one term association measure, Z-score, which showed good performance in query expansion experiments (Vechtomova et al., 2003). 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 very similar to a t-score (Church et al., 1994), the difference is that Z is used for the data distributed normally.

We used the method for extracting collocates and calculating Z-score as proposed in (Vechtomova et al., 2003). 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. 7 (Vechtomova et al., 2003), 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.

$$Z = \frac{f_r(x, y) - \frac{f_c(y) f_r(x) v_x(R)}{N}}{\sqrt{\frac{f_c(y) f_r(x) v_x(R)}{N}}} \qquad (7)$$

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 (Vechtomova et al., 2003). 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 (Vechtomova et al., 2003). 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 chapter. The full list of selected collocates for all queries submitted by users is listed in Appendix A.

Bill Gates	wealth, dynamite, Microsoft, rich, interview, Napoleon, dollars, he, man, person, short, say
Dandy Warhols	lyrics, warlords, odditorium, mars, smoke, rave, tabs, dig, driving, tour, artists, music
Egypt	ancient, guardian, pyramids, tour, arab, egyptian, travel, Nile, Cairo, modern, country, history
Firefly	monologue, serenity, episodes, series, Whedon, Joss, inner, television, episode, show, official, TV
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
softwood lumber dispute	Canada, BC, trade, canadian, US

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

3.5.2 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 $p(d_i)$ from precomputed statistics (section 3.3) that it refers to a query term or a collocate instance *i* occurring *d* words away. Aggregate probabilities are calculated according to Eq. 5 (section 3.4).

Chapter 4

Evaluation

4.1 Evaluation in IR

In Information Retrieval strong emphasis has always been placed on unbiased and rigorous evaluation of developed methods. Large-scale evaluation is frequently done using test collections which consist of user queries (also referred to as topics), a collection of documents and user relevance judgements, specifying whether a document is relevant to the query or not. The large-scale experimentation tradition of using test collections began with Cranfield tests in the 50s and 60s (Cleverdon, 1991), and has been continued for the past 15 years by the Text Retrieval Conferences (TREC) (Voorhees and Buckland 2004), and more recently by the Cross Language Evaluation Forum (CLEF) (Peters et al, 2005).

Text Retrieval Conference is co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense. The main purpose of TREC is to provide infrastructure for large-scale evaluation of text retrieval methodologies, and to provide an open forum for IR researchers. TREC consists of several tracks, or specific IR tasks, such as Ad-hoc, Question Answering, High Accuracy Retrieval from Documents, Enterprise, etc. TREC has an annual cycle, in which the participating research groups develop a system for a particular track, receive document collections and a set of user queries (topics) from track organisers at NIST, use these queries to retrieve a set of documents (or other information items, such as passages or text snippets) and return them to NIST. Evaluators hired by NIST judge the relevance of the top retrieved documents in the submitted runs.

Traditionally relevance has been considered as a binary concept, i.e. a document is either relevant or non-relevant to a query, however, in recent years, some research has been done on judging document relevance according to several degrees of relevance. In our evaluation, as will be described in Section 4.3 we used two types of relevance judgements "Query Relevance" (i.e. the document contains opinion on the query topic) and "Related Topic Relevance" (i.e. the document contains opinion on the topic related to the query topic).

After the relevance judgements have been made, the performance of each run is evaluated according to standard measures. The most frequently used measures are Mean Average Precision, R-Precision and Precision at various document cutoff points, such as Precision at 10 retrieved documents (P@10) and Precision at 15 documents (P@15).

Mean Average Precision is calculated over a set of topics from the Average Precision values for each topic. Average Precision is the average of the precision after each relevant document is retrieved and is calculated according to Eq. 8.

$$AveP = \frac{\sum_{r=1}^{N} Prec(r)}{N} \qquad (8)$$

Where, *N* – number of relevant documents retrieved for the topic; *r* – rank of the retrieved relevant document; *Prec*(*r*) – precision at rank *r* calculated as in Eq. 9:

$$Prec(r) = \frac{n(r)}{r} \qquad (9)$$

Where, *n*(*r*) – number of relevant documents retrieved up to and including rank *r*.

R-Precision is precision after *R* documents have been retrieved where *R* is the number of relevant documents for the topic.

Precision at *x* document cutoff point (e.g., P@10 and P@15) is calculated according to Eq 9.

In addition to laboratory type of evaluations using test collections and judgements by trained evaluators, there is also a user-based evaluation methodology, which relies on the participation of real users in the experiments. Such evaluation methodology is particularly suitable for the tasks for which no appropriate test collection exists, or if the object of evaluation is an interactive system. Such evaluation methods may yield richer and more realistic data about the system and its use by real users with genuine information needs in real-life settings. However, such experiments need to be carefully controlled in order to avoid the effect of nuisance variables, such as the order in which the control and experimental systems are presented to the users. Also, it is usually possible only to evaluate a limited number of system variables, as the number of available participants and their time is limited and therefore each user can perform only a small number of interactions with the system. User-based experiments typically require more time to conduct and can be more expensive than laboratory experiments with test collections, as time is needed for recruiting sufficient number of participants, and collecting data from them. In this work we have chosen to evaluate the developed methods by means of a user-based evaluation methodology. The motivation for this choice is presented in the next section, while the detailed evaluation methodology we followed is given in section 4.3.

4.2 Selection of the Evaluation Methodology

A number of candidate test collections were considered for evaluating our proposed method, however, none of them seemed suitable for our task. One of the candidates considered was the test collection from the HARD (High Accuracy Retrieval from Documents) track of TREC 2004 (Allan, 2005). The corpus used in the HARD 2004 track was a collection of newswire articles of either opinion/editorial or news report genre. User queries (topics) had additional metadata, specifying among other criteria whether the user is searching for opinion/editorial articles or news reports. Each document judged by the user was marked as either "off-topic" (not relevant to the query topic), "on-topic" (relevant to the query topic but not satisfying all metadata criteria), and "relevant" (relevant to the query topic and satisfying all query metadata criteria). One problem with using this test collection was a very small number of topics requesting opinion/editorial articles (10 topics out of 50). However, a bigger problem was that although a document may be of opinion/editorial genre and relevant to the query topic, it may not express any direct opinion on the concept expressed by the query. Also, topics used in the HARD track typically do not represent a specific opinion target, but rather a broader subject about which opinion/editorial articles are sought.

Another possible candidate test collection is Enterprise Track 2005 discussion search task (Craswell et al., 2005). The document collection used in this track consists of 198,394 emails from email lists of W3C. The goal of the task was to find email messages contributing to a discussion on the topic of the query and containing either a pro or a con argument regarding the topic. Queries were formulated by participating groups and each document was judged as non-relevant, partially relevant (i.e. relevant without a pro/con statement) and relevant (with a pro/con statement). This collection became available after we have commenced our evaluation experiment. Also in our opinion it is not suitable for our purposes as it is domain-specific, and not all topics lend themselves to pro/con statements on the subject.

We also considered domain-specific corpora, such as film reviews. Although film reviews have a clear target – the film title, the problem is that only opinionated documents are available.

Lacking a domain-independent test collection of queries and documents judged as having or not having an opinion on the concept represented by the query, we decided to conduct a user-based evaluation using the Web as corpus. By means of such evaluation methodology we can test the proposed approach in an unrestricted domain with queries representing genuine information needs of users. It can also give us an insight into the kind of entities users want to find opinions on. The users' queries and judgements may also be useful as a test collection for other researchers.

4.3 Evaluation Methodology

Altogether 33 users, solicited from the University of Waterloo graduate student mailing list, voluntarily participated in the evaluation. The user-based evaluation experiment design was reviewed by the University of Waterloo Ethics Committee and received full ethics clearance in March 2005. The Letter of Notification of Full Ethics Clearance is given in Appendix D. The User Recruitment Letter is given in Appendix E, Information Letter in Appendix F, Survey Questions in Appendix G, and Feedback Page in Appendix H. The evaluation was conducted over a period of 10 months (June 2005 – March 2006).

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. The documents used for re-ranking consist of all results obtainable via the Google API up to the first thousand. These results are returned, similar to Google pages, in groups of ten per request. Oftentimes, instead of ceasing to provide results, Google will instead return duplicate pages. Once a result set contains no hitherto unseen URLs it is estimated that no additional results are forthcoming and the search is terminated. Only documents in English were requested, however, specifying English as the language of the documents has been observed to still permit, on occasion, many non-English results. To reduce the impact of this, only documents containing at least two of the top five highest frequency English words in the British National Corpus with more than two characters[5] are selected.

We evaluated two methods of ranking documents by topical opinion:

1) "Opinion" method using only original query terms (presented in section 3.4);
2) "Collocation opinion" method using original query terms plus their collocates (presented in section 3.5.2).

The baseline against which the above two methods are evaluated is the original Google ranking. 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.

[5] *"the"*, *"and"*, *"was"*, *"with"*, *"that"*, as documented at ftp://ftp.itri.bton.ac.uk/bnc/written.num.o5 (Accessed: April 2006)

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.

Each user submitted one query, and in total for 33 queries 1192 documents were judged.

Chapter 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 averaged over 33 queries are presented in Table 3, while the results for individual queries are given in Appendix B.

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*

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

As can be seen from Table 3 all runs significantly (paired t-test, P<0.05) improved performance of topical opinion retrieval over the Google baseline. The results of the t-test are given in Appendix C. 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 queries (Fig. 1). In P@10 "opinion" method improves performance in 21 cases, deteriorates 6, and does not affect performance in 6 cases, while "collocation opinion" improves the performance of 20 queries, deteriorates 10, and does not affect the performance of 3 queries (Fig. 2).

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 (Fig. 3). In P@10 "opinion" method improves performance in 22 queries, deteriorates 6, and does not affect performance of 5 queries. "Collocation opinion" also improves P@10 of 22 queries, but deteriorates 11 queries (Fig. 4).

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 s/he may be interested in documents expressing opinions on subjects related to their query topic.

The difference between the two opinion scoring methods is not statistically significant in any measure using paired t-test at $P<0.05$ (Appendix C). It is likely that collocates introduce some degree of noise to the opinion scoring process: for example if a query is "Microsoft" and the collocate is "company", it is possible that the latter refers to some company other than Microsoft in a document. There is a considerable scope for improving the collocation-based method by using different collocate ranking methods and tuning collocate selection parameters. Also, it makes sense to downplay the effect of collocates on the document score compared to the original query terms.

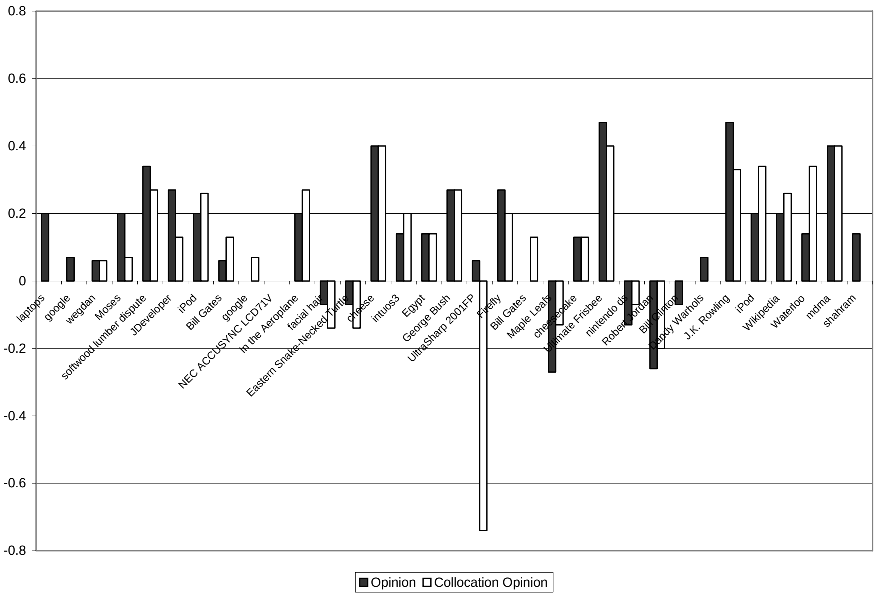


Figure 1. Precision at 15 (P@15) results for Opinion and Collocation Opinion ("Query Relevance" judgements).

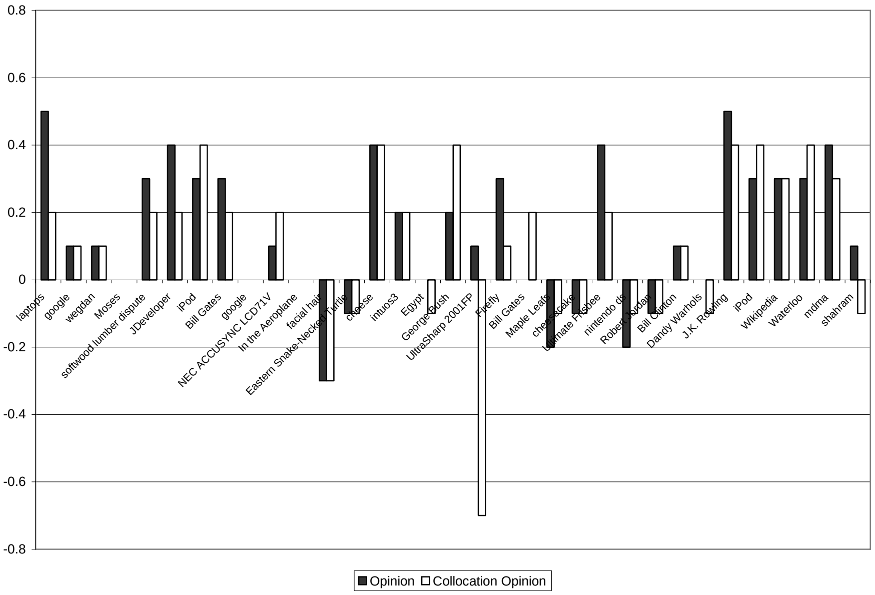


Figure 2. Precision at 10 (P@10) results for Opinion and Collocation Opinion ("Query Relevance" judgements).

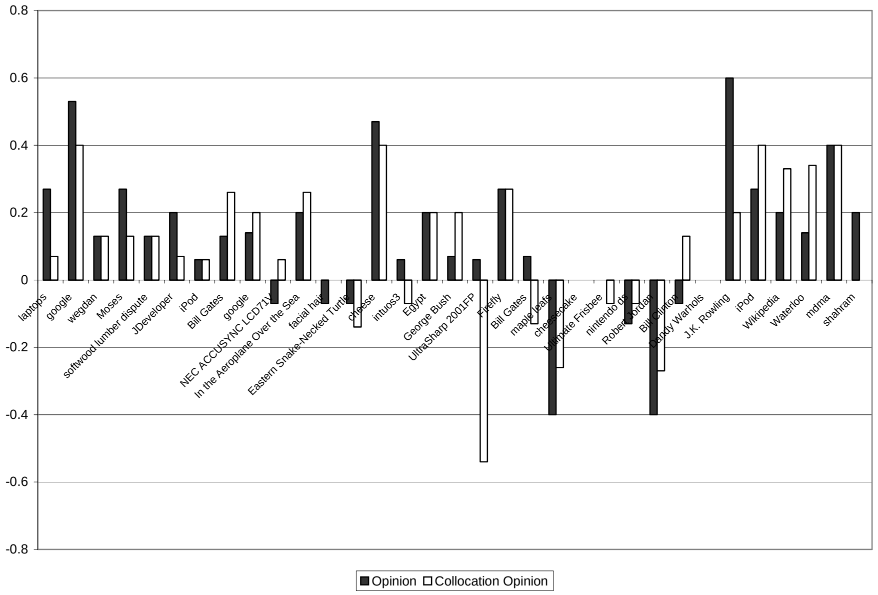


Figure 3. Precision at 15 (P@15) results for Opinion and Collocation Opinion ("Related Topic Relevance" judgements).

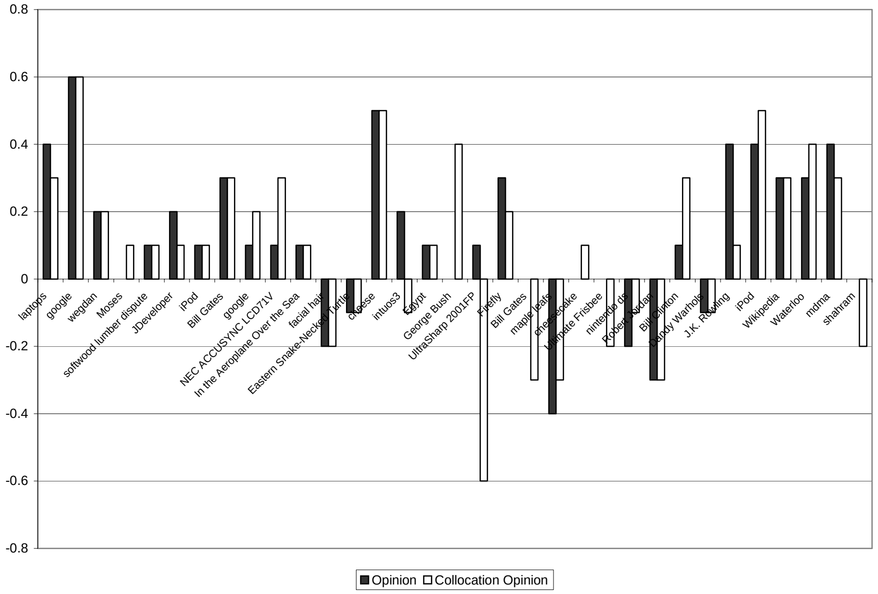


Figure 4. Precision at 10 (P@10) results for Opinion and Collocation Opinion ("Related Topic Relevance" judgements).

Chapter 6

Conclusions And Future Work

This thesis reports a novel method for topical opinion retrieval, and demonstrates the usefulness of the developed approach through a user-based evaluation showing a significant result. The algorithm is computationally lightweight and lends itself to implementation in a system for ad hoc document retrieval. As an element of our technique, we developed a method for adjective target resolution in English.

There is room for improvement in the statistics used to determine the likelihood of a subjective reference to the query. For training purposes a newswire rather than web corpus was used, because the web data is not always structured into sentences and contains navigational and other non-prose elements, which increase the complexity of the parsing task. These same elements also skew the ultimate application of the statistics. Tools for elimination of navigational and template elements in hypertext hold promise for larger gains in precision as the calculated statistics would more closely represent the reality it is used to predict.

We did not explore the many options available to select collocates. Use of the Z-score with different parameters or other collocate selection measures such as Log-Likelihood or language modelling approaches might lead to further improvements. It may be beneficial to constrain the process to collocates from windows containing subjective adjectives to specifically target features or attributes of the query that are evaluated. Excluding terms found immediately adjacent to query terms may eliminate idiomatic and syntactic relationships. Rather than considering collocate terms to have the same importance as the query, weighting schemes could be investigated. One that comes to mind would be to incorporate the relevance ranking of the document or passage as more topical text seems more likely to contain topical collocates.

Rather than simply using a list of subjective adjectives, the system could incorporate additional metainformation on each term 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.

Another opportunity for expressiveness would be to allow queries to combine relevance and opinion matching. One example would be searching for pages about a BMW M3 with opinions on transmission. Brief experimentation with this has produced interesting yet still anecdotal results.

Appendix A

Collocates of User Query Terms

Query	**Collocates**
Bill Gates	wealth, dynamite, microsoft, rich, interview, napoleon, dollars, he, man, person, short, say
Bill Gates	wealth, dynamite, microsoft, rich, interview, napoleon, dollars, he, man, person, short, say
Dandy Warhols	lyrics, warlords, odditorium, mars, smoke, rave, tabs, dig, driving, tour, artists, music
Eastern Snake-Necked Turtle	longicollis, chelodina
Egypt	ancient, guardian, pyramids, tour, arab, egyptian, travel, nile, cairo, modern, country, history
Firefly	monologue, serenity, episodes, series, whedon, joss, inner, television, episode, show, official, tv
George Bush	funny, library, pictures, president, bush
In the Aeroplane Over the Sea	neutral, milk, hotel, lyrics, album, two, music
J.K. Rowling	bloomsbury, potter, harry, author, interview, book, books, site
Jdeveloper	oracle, 10g, soa, oc4j, webgalileo, oracle9i, ide, bpel, clover, jsf, adf, java
Moses	miriam, cone, monotheism, pharaoh, aaron, mendelssohn, exodus, akhenaten, grandma, torah, hebrew, birth
NEC	monitor, tag, driver, flat, lcd, panel, value, inch, nec, price

ACCUSYNC LCD71V	
Robert Jordan	fandom, wheel, series, time, hardcover, books, author, world, book, page, posted, he
Ultimate Frisbee	disc, ultimate
UltraSharp 2001FP	dell, lcd, monitor, flat, tft, panel, gray, brilliant, matrix, sharp, displays, inch
Wikipedia	deletion, log, votes, carey, evans, encyclopedia, wiki, judge, january, debates, february, articles
Bill Clinton	jokes, president, presidency, former, life, clinton, house, white, bush, he
cheese	llangloffan, farmhouse, cheeses, camembert, vermont, fondue, jenkins, ripening, pairing, diaries, stilton, milk
cheesecake	chocolate, extraordinaire, pumpkin, coconut, praline, recipes, strawberry, almond, pineapple, raspberry, bake, cherry
facial hair	removal, trimmer, beards, women, philips, remove, growth, men, hair, cream, style, can
Google	philipp, quixtar, lenssen, montage, orkut, ipo, search, bombing, toolbar, analysts, bomb, cache
Google	philipp, quixtar, lenssen, montage, orkut, ipo, search, bombing, toolbar, analysts, bomb, cache
iPod	nano, apple, itunes, grayscale, 30gb, generation, mini, dock, gb, shuffle, applecare, playback
iPod	nano, apple, itunes, grayscale, libipoddevice, 30gb, generation, mini, dock, shuffle, gb, dockable
Intuos3	wacom, pen, tablet, 6x8, 9x12, graphire4, grip, 4x5, expresskeys, usb, mouse, 21ux
laptops	sager, notebooks, notebook, laptop, discount, computers, cheap, centrino, star, pro, specializing, p4
Maple Leafs	toronto, totonto, tickets, vs, lindros, nhl, canadiens, capitals, thrashers, hockey, hurricanes, tampa
mdma	ptsd, psychotherapy, serotonin, ecstasy, neurotoxicity, doses, effects, analogues, reuptake, ht1a, shulgin, acetylcholine
Nintendo ds	game, nintendogs, console, hardware, region, protection, games, ds, advance, nintendo, gba, boy
Shahram	kholdi, latifi, ghandeharizadeh, shabpareh, pradip, bagherzadeh, ee, nazeri, izadi, entekhabi, nader, posted
softwood lumber dispute	canada, bc, trade, canadian, us
Waterloo	kitchener, hawks, richmond, region, london, inn, battle, ontario, university, welcome, city, black
Wegdan	abdelsalam

Table A1: Collocates of users' query terms selected using Z score statistic and used in the "collocation opinion" document ranking method.

Appendix B

Evaluation results

Topic	P@10			P@15		
	Baseline	Opinion	Collocation Opinion	Baseline	Opinion	Collocation Opinion
laptops	0.1	0.6	0.3	0.33	0.53	0.33
Google	0.1	0.2	0.2	0.13	0.20	0.13
wegdan	0.1	0.2	0.2	0.07	0.13	0.13
Moses	0.5	0.5	0.5	0.40	0.60	0.47
softwood lumber dispute	0.5	0.8	0.7	0.53	0.87	0.80
JDeveloper	0.3	0.7	0.5	0.40	0.67	0.53
iPod	0.6	0.9	1	0.67	0.87	0.93
Bill Gates	0.6	0.9	0.8	0.67	0.73	0.80
Google	0	0	0	0.00	0.00	0.07
NEC ACCUSYNC LCD71V	0.1	0.2	0.3	0.20	0.20	0.20
In the Aeroplane over the Sea	0.8	0.8	0.8	0.60	0.80	0.87
facial hair	0.6	0.3	0.3	0.47	0.40	0.33
Eastern Snake-Necked Turtle	0.6	0.5	0.5	0.47	0.40	0.33
cheese	0.4	0.8	0.8	0.40	0.80	0.80
intuos3	0.1	0.3	0.3	0.13	0.27	0.33
Egypt	0.5	0.5	0.4	0.33	0.47	0.47
George Bush	0.2	0.4	0.6	0.33	0.60	0.60
UltraSharp 2001FP	0.9	1	0.2	0.87	0.93	0.13
Firefly	0.6	0.9	0.7	0.53	0.80	0.73
Bill Gates	0.1	0.1	0.3	0.07	0.07	0.20
Maple Leafs	0.3	0.1	0.2	0.40	0.13	0.27
cheesecake	0.5	0.4	0.4	0.40	0.53	0.53
Ultimate Frisbee	0.4	0.8	0.6	0.33	0.80	0.73
nintendo ds	0.6	0.4	0.5	0.60	0.47	0.53
Robert Jordan	0.5	0.4	0.4	0.53	0.27	0.33
Bill Clinton	0.4	0.5	0.5	0.47	0.40	0.47
Dandy Warhols	0.3	0.3	0.2	0.33	0.40	0.33
J.K. Rowling	0	0.5	0.4	0.00	0.47	0.33
iPod	0.5	0.8	0.9	0.53	0.73	0.87
Wikipedia	0.4	0.7	0.7	0.27	0.47	0.53
Waterloo	0.1	0.4	0.5	0.13	0.27	0.47
mdma	0.6	1	0.9	0.53	0.93	0.93
shahram	0.1	0.2	0	0.13	0.27	0.13

Table A2: Precision at 10 (P@10) and Precision at 15 (P@15) using "Query Relevance" judgements.

Topic	P@10			P@15		
	Baseline	Opinion	Collocation Opinion	Baseline	Opinion	Collocation Opinion
laptops	0.2	0.6	0.5	0.40	0.67	0.47
Google	0.4	1	1	0.47	1.00	0.87
wegdan	0.1	0.3	0.3	0.07	0.20	0.20
Moses	0.5	0.5	0.6	0.40	0.67	0.53
softwood lumber dispute	0.8	0.9	0.9	0.80	0.93	0.93
JDeveloper	0.5	0.7	0.6	0.53	0.73	0.60
iPod	0.9	1	1	0.87	0.93	0.93
Bill Gates	0.6	0.9	0.9	0.67	0.80	0.93
Google	0.5	0.6	0.7	0.53	0.67	0.73
NEC ACCUSYNC LCD71V	0.1	0.2	0.4	0.27	0.20	0.33
In the Aeroplane over the Sea	0.8	0.9	0.9	0.67	0.87	0.93
facial hair	0.7	0.5	0.5	0.60	0.53	0.60
Eastern Snake-Necked Turtle	0.6	0.5	0.5	0.47	0.40	0.33
cheese	0.5	1	1	0.53	1.00	0.93
intuos3	0.8	1	0.7	0.87	0.93	0.80
Egypt	0.7	0.8	0.8	0.67	0.87	0.87
George Bush	0.4	0.4	0.8	0.53	0.60	0.73
UltraSharp 2001FP	0.9	1	0.3	0.87	0.93	0.33
Firefly	0.6	0.9	0.8	0.53	0.80	0.80
Bill Gates	0.8	0.8	0.5	0.73	0.80	0.60
maple leafs	0.5	0.1	0.2	0.53	0.13	0.27
cheesecake	0.6	0.6	0.7	0.73	0.73	0.73
Ultimate Frisbee	0.8	0.8	0.6	0.80	0.80	0.73
nintendo ds	0.6	0.4	0.5	0.60	0.47	0.53
Robert Jordan	1	0.7	0.7	1.00	0.60	0.73
Bill Clinton	0.5	0.6	0.8	0.60	0.53	0.73
Dandy Warhols	0.4	0.3	0.3	0.40	0.40	0.40
J.K. Rowling	0.3	0.7	0.4	0.20	0.80	0.40
iPod	0.5	0.9	1	0.53	0.80	0.93
Wikipedia	0.4	0.7	0.7	0.27	0.47	0.60
Waterloo	0.1	0.4	0.5	0.13	0.27	0.47
mdma	0.6	1	0.9	0.53	0.93	0.93
shahram	0.2	0.2	0	0.20	0.40	0.20

Table A3: Precision at 10 (P@10) and Precision at 15 (P@15) using “Relevance to a related topic” judgements.

Appendix C

T-test results

	Baseline	*Opinion*
Mean	0.376	0.518
Variance	0.058	0.082
Observations	33	33
Pearson Correlation	0.687	
Hypothesized Mean Difference	0.000	
df	32	
t Stat	-3.856	
P(T<=t) one-tail	0.000	
t Critical one-tail	1.694	
P(T<=t) two-tail	0.001	
t Critical two-tail	2.037	

Table A4: Paired t-test for the P@10 results (Query Relevance judgements) for the *Baseline* and *Opinion* runs.

	Baseline	*Collocation Opinion*
Mean	0.376	0.473
Variance	0.058	0.067
Observations	33	33
Pearson Correlation	0.538	
Hypothesized Mean Difference	0.000	
df	32	
t Stat	-2.317	
P(T<=t) one-tail	0.014	
t Critical one-tail	1.694	
P(T<=t) two-tail	0.027	
t Critical two-tail	2.037	

Table A5: Paired t-test for the P@10 results (Query Relevance judgements) for the *Baseline* and *Collocation Opinion* runs.

	Opinion	*Collocation Opinion*
Mean	0.518	0.473
Variance	0.082	0.067
Observations	33	33
Pearson Correlation	0.789	
Hypothesized Mean Difference	0.000	
df	32	
t Stat	1.461	
P(T<=t) one-tail	0.077	
t Critical one-tail	1.694	
P(T<=t) two-tail	0.154	
t Critical two-tail	2.037	

Table A6: Paired t-test for the P@10 results (Query Relevance judgements) for the *Opinion* and *Collocation Opinion* runs.

	Baseline	*Opinion*
Mean	0.372	0.499
Variance	0.044	0.072
Observations	33	33
Pearson Correlation	0.735	
Hypothesized Mean Difference	0.000	
df	32	
t Stat	-3.998	
P(T<=t) one-tail	0.000	
t Critical one-tail	1.694	
P(T<=t) two-tail	0.000	
t Critical two-tail	2.037	

Table A7: Paired t-test for the P@15 results (Query Relevance judgements) for the *Baseline* and *Opinion* runs.

	Baseline	*Collocation Opinion*
Mean	0.372	0.475
Variance	0.044	0.068
Observations	33	33
Pearson Correlation	0.558	
Hypothesized Mean Difference	0.000	
df	32	
t Stat	-2.617	
P(T<=t) one-tail	0.007	
t Critical one-tail	1.694	
P(T<=t) two-tail	0.013	
t Critical two-tail	2.037	

Table A8: Paired t-test for the P@15 results (Query Relevance judgements) for the *Baseline* and *Collocation Opinion* runs.

	Opinion	*Collocation Opinion*
Mean	0.499	0.475
Variance	0.072	0.068
Observations	33	33
Pearson Correlation	0.801	
Hypothesized Mean Difference	0.000	
df	32	
t Stat	0.832	
P(T<=t) one-tail	0.206	
t Critical one-tail	1.694	
P(T<=t) two-tail	0.411	
t Critical two-tail	2.037	

Table A9: Paired t-test for the P@15 results (Query Relevance judgements) for the *Opinion* and *Collocation Opinion* runs.

	Baseline	*Opinion*
Mean	0.542	0.664
Variance	0.058	0.073
Observations	33	33
Pearson Correlation	0.603	
Hypothesized Mean Difference	0.000	
df	32	
t Stat	-3.043	
P(T<=t) one-tail	0.002	
t Critical one-tail	1.694	
P(T<=t) two-tail	0.005	
t Critical two-tail	2.037	

Table A10: Paired t-test for the P@10 results (Relevance to a related topic judgements) for the *Baseline* and *Opinion* runs.

	Baseline	*Collocation Opinion*
Mean	0.542	0.636
Variance	0.058	0.066
Observations	33	33
Pearson Correlation	0.386	
Hypothesized Mean Difference	0.000	
df	32	
t Stat	-1.963	
P(T<=t) one-tail	0.029	
t Critical one-tail	1.694	
P(T<=t) two-tail	0.058	
t Critical two-tail	2.037	

Table A11: Paired t-test for the P@10 results (Relevance to a related topic judgements) for the *Baseline* and *Collocation Opinion* runs.

	Opinion	*Collocation Opinion*
Mean	0.664	0.636
Variance	0.073	0.066
Observations	33	33
Pearson Correlation	0.738	
Hypothesized Mean Difference	0.000	
df	32	
t Stat	0.821	
P(T<=t) one-tail	0.209	
t Critical one-tail	1.694	
P(T<=t) two-tail	0.418	
t Critical two-tail	2.037	

Table A12: Paired t-test for the P@10 results (Relevance to a related topic judgements) for the *Opinion* and *Collocation Opinion* runs.

	Baseline	*Opinion*
Mean	0.545	0.663
Variance	0.051	0.062
Observations	33	33
Pearson Correlation	0.584	
Hypothesized Mean Difference	0.000	
df	32	
t Stat	-3.088	
P(T<=t) one-tail	0.002	
t Critical one-tail	1.694	
P(T<=t) two-tail	0.004	
t Critical two-tail	2.037	

Table A13: Paired t-test for the P@15 results (Relevance to a related topic judgements) for the *Baseline* and *Opinion* runs.

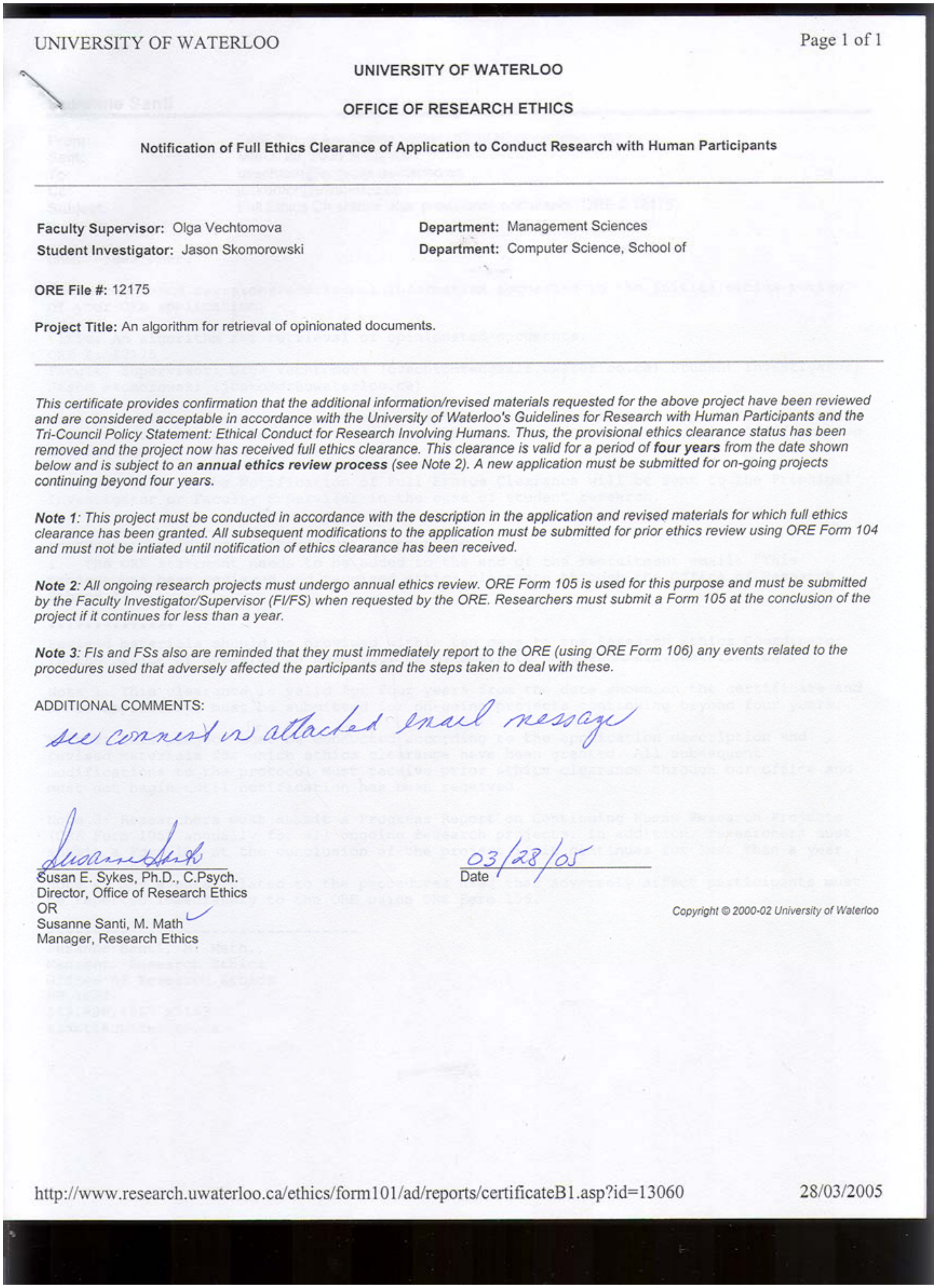
	Baseline	*Collocation Opinion*
Mean	0.545	0.640
Variance	0.051	0.057
Observations	33	33
Pearson Correlation	0.567	
Hypothesized Mean Difference	0.000	
df	32	
t Stat	-2.517	
P(T<=t) one-tail	0.009	
t Critical one-tail	1.694	
P(T<=t) two-tail	0.017	
t Critical two-tail	2.037	

Table A14: Paired t-test for the P@15 results (Relevance to a related topic judgements) for the *Baseline* and *Collocation Opinion* runs.

	Opinion	*Collocation Opinion*
Mean	0.663	0.640
Variance	0.062	0.057
Observations	33	33
Pearson Correlation	0.757	
Hypothesized Mean Difference	0.000	
df	32	
t Stat	0.749	
P(T<=t) one-tail	0.230	
t Critical one-tail	1.694	
P(T<=t) two-tail	0.460	
t Critical two-tail	2.037	

Table A15: Paired t-test for the P@15 results (Relevance to a related topic judgements) for the *Opinion* and *Collocation Opinion* runs.

Appendix D
Research Ethics Clearance



UNIVERSITY OF WATERLOO Page 1 of 1

UNIVERSITY OF WATERLOO

OFFICE OF RESEARCH ETHICS

Notification of Full Ethics Clearance of Application to Conduct Research with Human Participants

Faculty Supervisor: Olga Vechtomova **Department:** Management Sciences
Student Investigator: Jason Skomorowski **Department:** Computer Science, School of

ORE File #: 12175

Project Title: An algorithm for retrieval of opinionated documents.

This certificate provides confirmation that the additional information/revised materials requested for the above project have been reviewed and are considered acceptable in accordance with the University of Waterloo's Guidelines for Research with Human Participants and the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans. Thus, the provisional ethics clearance status has been removed and the project now has received full ethics clearance. This clearance is valid for a period of ***four years*** *from the date shown below and is subject to an* ***annual ethics review process*** *(see Note 2). A new application must be submitted for on-going projects continuing beyond four years.*

Note 1: *This project must be conducted in accordance with the description in the application and revised materials for which full ethics clearance has been granted. All subsequent modifications to the application must be submitted for prior ethics review using ORE Form 104 and must not be intiated until notification of ethics clearance has been received.*

Note 2: *All ongoing research projects must undergo annual ethics review. ORE Form 105 is used for this purpose and must be submitted by the Faculty Investigator/Supervisor (FI/FS) when requested by the ORE. Researchers must submit a Form 105 at the conclusion of the project if it continues for less than a year.*

Note 3: *FIs and FSs also are reminded that they must immediately report to the ORE (using ORE Form 106) any events related to the procedures used that adversely affected the participants and the steps taken to deal with these.*

ADDITIONAL COMMENTS:

see comment in attached email message

Susan E. Sykes, Ph.D., C.Psych.
Director, Office of Research Ethics
OR
Susanne Santi, M. Math ✓
Manager, Research Ethics

03/28/05
Date

Copyright © 2000-02 University of Waterloo

http://www.research.uwaterloo.ca/ethics/form101/ad/reports/certificateB1.asp?id=13060 28/03/2005

Appendix E

User Recruitment Email

Fellow Graduate Students:

I require your help to evaluate the information retrieval method I'm developing for my thesis. It focuses on the retrieval of documents that have an opinion about your query. All I require is a few minutes of your time, less than an hour in all, to help me decide if it actually works by providing a query then returning the next day to label which resulting documents actually express an opinion on that topic.

If this sounds like something you might be willing to lend a hand with, the query form along with additional detail on the project is available at <URL of Information Letter> .

Thanks,

Jason Skomorowski
MMath Student
School of Computer Science
University of Waterloo
Email: jcskomor@uwaterloo.ca

Appendix F

Informational Letter

<displayed upon initiation of survey>

Dear Participant,

The project is part of the MMath thesis of Jason Skomorowski, conducted under supervision of Dr. Olga Vechtomova (Assistant Professor, Department of Management Sciences, University of Waterloo, Email: ovechtom@engmail.uwaterloo.ca, Phone: 519 888 4567 ext. 2675).

The goal of the project is to develop an Information Retrieval algorithm for the effective retrieval and ranking of opinionated documents. Users of search engines frequently search for documents, which express opinion about a certain subject, for example product/music/book reviews. Current search algorithms do not differentiate between documents which treat the subject objectively and those which contain opinions about them. Our objective is to develop a document ranking algorithm to rank highly those documents which express opinion about the subject of the searcher's query. The main research question is whether the proposed algorithm retrieves and ranks opinionated documents more effectively for this purpose than the traditional document retrieval algorithms.

Your participation will require less than an hour of time in two separate sessions and involve the submission of a query and the evaluation of resulting documents for level of opinion. While participants will not benefit personally from completing this entirely optional survey, we hope it will lead to better methods of document retrieval. At no time will identifying information be collected from you – the survey is anonymous. If you have any questions about this research, please do not hesitate to email jcskomor@uwaterloo.ca

This project has been reviewed by, and received ethics clearance through, the Office of Research Ethics. In the event you have any comments or concerns resulting from your participation in this study, please contact Dr. Susan Sykes at 519-888-4567, Ext. 6005.

Sincerely Yours,
Jason Skomorowski
jcskomor@uwaterloo.ca

Appendix G

Survey Questions

The survey is in two parts:

Part A: Query

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

<text input field>

Describe below in a few sentences what you are seeking with this query so we can better interpret the results.

<text input field>

Part B: Results

Does the following document:
() Express an opinion about the query item.
() Express an opinion on something related to the query item (e.g., an opinion of golf clubs when queried on golf)
() Express no opinion regarding the query.

<body of document with query terms highlighted>

Appendix H

Feedback Page

<displayed upon completion of survey>

Dear Participant,

We would like to thank you for your time and commitment to this study. It is very valuable to our project.

The data collected during interviews will contribute to a better understanding of how to more effectively retrieve opinionated documents.

A summary of the main findings of the project will be made available to all participants upon request. Please email jcskomor@uwaterloo.ca if you are interested in receiving this information as soon as the research is complete.

Sincerely,

Jason Skomorowski
MMath Student
School of Computer Science
University of Waterloo
Email: jcskomor@uwaterloo.ca

Bibliography

J. Allan, HARD Track overview in TREC 2004. High Accuracy Retrieval From Documents. In Voorhees, E. and Buckland, L. (Eds.) TREC 2004 Proceedings, Gaithesburg, MD, 2005.

M. C. Baker, Lexical categories: verbs, nouns and adjectives. Cambridge University Press, 2003.

S. Bethard, H. Yu, A. Thornton, V. Hatzivassiloglou, and D. Jurafsky, Automatic extraction of opinion propositions and their holders. In Working Notes of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, 2004.

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.

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.

N. Craswell, A. de Vries and I. Soboroff, Overview of the TREC-2005 Enterprise Track. In E. Voorhees, L. Buckland (Eds.) Proceedings of the 14th Text Retrieval Conference, Gaithersburg, 2005.

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.

R. M. W. Dixon and A. Y. Aikhenvald, Adjective classes. a crosslinguistic typology. Oxford University Press, 2004.

A. Esuli and F. Sebastiani Determining the Semantic Orientation of Terms through Gloss Classification. In Proceedings of CIKM 2005, Bremen, Germany, pp. 617-624.

S. Greenbaum, The Oxford English grammar. Oxford University Press, 1996.

D. J. Harper, C. J. Van Rijsbergen, An evaluation of feedback in document retrieval using co-occurrence data. Journal of Documentation, 34(3), 1978, pp.189-216.

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.

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.

M. Hurst and K. Nigam, Retrieving topical sentiments from online document collections. In proceedings of the 11th conference on document recognition and retrieval, 2004.

S. Kim and E. Hovy Identifying and Analyzing Judgment Opinions. In Proceedings of HLT-NAACL 2006, New York, US, 2006.

X. Li and D. Roth, Exploring evidence for shallow parsing. In Proceedings of the Annual Conference on Computational Natural Language Learning, 2001.

D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, Massachusetts, 1999.

G.A. Miller Wordnet: an on-line lexical database. International Journal of Lexicography, 3(4), 235-312, 1990.

M. Munoz, V. Punyakanok, D. Roth, and D. Zimak, A Learning Approach to Shallow Parsing. EMNLP-VLC'99, the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999.

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.

C. Peters, P. Clough, J. Gonzalo, G.J.F. Jones, M. Kluck and B. Magnini (Eds.) Multilingual Information Access for Text, Speech and Images. Fifth Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004.

P. Rijkhoek, On Degree Phrases and Result Clauses. PhD thesis, University of Groningen, Groningen, 1998.

S.E. Robertson, S. Walker, S. Jones, M. Hankock-Beaulieu, M. Gatford, Okapi at TREC-3. In Harman D. (Ed.) Proceedings of the Third Text Retrieval Conference, NIST, Gaithersburg, U.S., pp.109-126, 1995.

J. Sinclair (Ed.) Collins Cobuild English Grammar. Harper Collins, 1990.

A.F. Smeaton and C. J. Van Rijsbergen, The retrieval effects of query expansion on a feedback document retrieval system. The Computer Journal, 26(3), 1983, pp. 239-246.

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

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.

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.

J. Van Rijsbergen, A theoretical basis for the use of co-occurrence data in information retrieval. Journal of Documentation: 33(2), 1977, pp. 106-119.

O. Vechtomova, S. E. Robertson, and S. Jones, Query expansion with long-span collocates. Information Retrieval, vol. 6, pp. 251-273, 2003.

Z. Vendler, Adjectives and nominalizations, p. 86. Mouton & Co. N.V., The Hague, 1968.

E. Voorhees and L. Buckland, Proceedings of the Twelfth Text Retrieval Conference, NIST, Gaithersburg, US, 2004.

C. Whitelaw, N. Garg and S. Argamon, Using Appraisal Groups for Sentiment Analysis. In Proceedings of CIKM 2005, Bremen, Germany, pp. 625-631.

J. Wiebe, Learning subjective adjectives from corpora. In Proceedings of the 17th National Conference on Artificial Intelligence, pp. 735-740, 2000.

J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, Learning subjective language. Computational Linguistics, vol. 30, no. 3, pp. 277-308, 2004.

J. Xu and B. Croft, Query expansion using local and global document analysis. In: the 19th ACM Conference on Research and Development in Information Retrieval (SIGIR '96), ACM Press, New York, 1996. pp. 4-11.

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.