Na, J.-C., Khoo, C., & Chan, S. (2006). A sentiment-based meta search engine. In C. Khoo, D. Singh & A.S. Chaudhry (Eds.), *Proceedings of the Asia-Pacific Conference on Library & Information Education & Practice 2006 (A-LIEP 2006), Singapore, 3-6 April 2006* (pp. 83-89). Singapore: School of Communication & Information, Nanyang Technological University.

A SENTIMENT-BASED META SEARCH ENGINE

JIN-CHEON NA & CHRISTOPHER KHOO

School of Communication & Information, Nanyang Technological University,
31 Nanyang Link, Singapore 637718
E-mail: {tjcna, assgkhoo}@ntu.edu.sg

SYIN CHAN

School of Computer Engineering, Nanyang Technological University,
N4-02c-72 Nanyang Avenue, Singapore 639798
E-mail: asschan@ntu.edu.sg

Abstract. This study is in the area of sentiment classification—classifying online review documents according to the overall sentiment expressed in them. This paper presents a prototype sentiment-based meta search engine that has been developed to perform sentiment categorization of Web search results. It assists users to quickly focus on *recommended* or *non-recommended* information by classifying Web search results into four categories: *positive*, *negative*, *neutral*, and *non-review* documents. It does this by using an automatic classifier based on a supervised machine learning algorithm, Support Vector Machine (SVM). This paper also discusses various issues we have encountered during the prototype development, and presents our approaches for resolving them. A user evaluation of the prototype was carried out with positive responses from users.

Introduction

To help users locate relevant and useful information on the World Wide Web, researchers have developed classification/clustering tools to categorize Web search results. Chen and Dumais (2000) designed a user interface that automatically groups Web search results into topical categories such as *automotive*, *local interest*. The tool devised by Zeng, He, Chen, Ma and Ma (2004) uses salient phrases extracted from the ranked list of documents as cluster names. For instance, with a query input, *Jaguar*, the generated cluster names are *Jaguar Cars*, *Panthera onca*, *Mac OS*, *Big Cats*, *Clubs*, and *Others*. Vivisimo (http://vivisimo.com) is an example of an operational clustering tool for Web search results. These tools, however, focused mainly on topical categorization—categorizing documents by subject or topical area.

In recent years, we have witnessed the tremendous growth of online discussion groups and review sites, where a crucial characteristic of the posted articles is their sentiment or overall opinion towards the subject matter. Researchers are turning their attention to a kind of non-topical classification called *sentiment classification*. Sentiment-related concepts can be characterized by at least three dimensions: type of sentiment/emotion/attitude, sentiment orientation (*positive* versus *negative*), and intensity of the sentiment. Our focus is on sentiment orientation—identifying a piece of document as having an overall positive or negative sentiment.

This paper introduces a sentiment-based meta search engine, based on *automatic sentiment classification*, which allows the user to specify a product name and subsequently categorizes the search results by the polarity of the desired reviews, such as *recommended* or *not recommended.* Our initial work was described in Na, Khoo, Chan and Hamzah (2005). It can help the user to focus on Web articles containing either positive or negative comments. For instance, a user who is mainly concerned about the negative aspects of a product can look at Web articles under the negative review category. The search results may also provide the overall distribution of positive and negative review pages, that is, the total numbers of positive and negative pages. In general, automatic sentiment classification is applicable to various types of materials stored in digital libraries. For instance, in a movie review digital library, it can be used to classify movie reviews collected by a Web crawler into positive and negative ones, so that potential moviegoers can have an overall idea of how a movie is perceived by other users. A sentiment timeline system can track online discussion about movies and display a plot of the number of positive and negative sentiment messages over time.

Section 2 discusses related works and our previous groundwork on sentiment classification. Sections 3 and 4 present our prototype system, a sentiment-based meta search engine. Section 5 describes the results of a preliminary user evaluation for the usability of the prototype system and a small ex-

periment to measure the sentiment classification accuracy of the system. Finally, Section 6 discusses future work and concludes the paper.

Related Works

Research in *automatic text classification* seeks to develop models for assigning category labels to new documents or document segments based on a set of training documents that have been classified by domain experts. Most studies of automatic text classification have focused on "topical classification", that is, classifying documents by subject or topic (e.g., *education* versus *entertainment*). Though machine-learning techniques have long been used in topical text classification with good results, they are less effective when applied to sentiment classification (Pang, Lee, & Vaithyanathan, 2002). Sentiment classification is a more difficult task compared to traditional topical classification, which classifies articles by comparing individual words (unigrams) in various subject areas.

Several researchers have carried out studies of automatic sentiment classification. Pang et al. (2002) examined the effectiveness of three machine learning methods (Naïve Bayes, Maximum Entropy, and Support Vector Machines SVM) for the sentiment classification of movie reviews (http://reviews.imdb.com/Reviews). They used mainly features based on unigrams (with negation tagging) and bigrams. SVM returned the best results (82.9% accuracy), using unigrams with binary weighting indicating the presence or absence of a feature.

Turney (2002) used an unsupervised machine learning technique to estimate the semantic orientation of a word based on its association with the words "excellent" and "poor", that is, the extent to which the word co-occurs with "excellent" and "poor" in a text collection. Mutual information was used as the association measure, and was computed using statistics (i.e., the number of hits returned) gathered by the Alta Vista search engine. The document phrases are bigrams, where one member of the pair is an adjective or an adverb and the second provides context. An average accuracy of 74% was achieved when the algorithm was evaluated on reviews from Epinions (http://www.epinions.com), sampled from four different domains (*automobiles*, *banks*, *movies*, and *travel destination*). The main limitation of this algorithm is the time required to calculate the semantic orientation of document phrases, done by sending queries to a search engine.

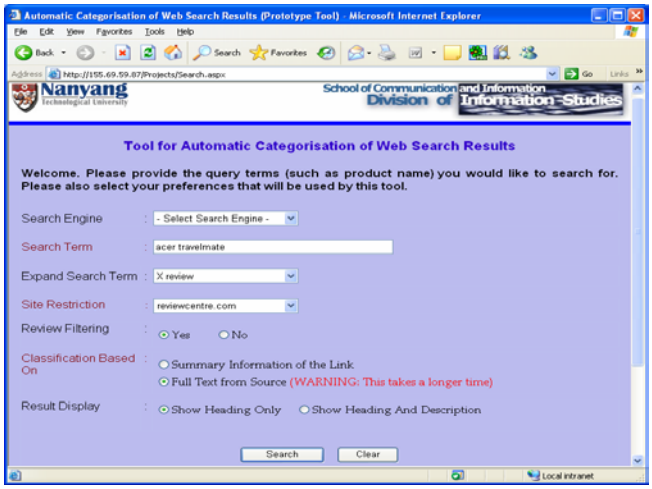
In a previous study (Na, Sui, Khoo, Chan, & Zhou, 2004), we investigated the effectiveness of using a machine learning algorithm (Support Vector Machine SVM) on various text features to classify online product reviews from Review Centre (www.reviewcentre.com) into *recommended* and *not recommended*. The majority of reviews were of mobile phones and electronic equipment. In sentiment classification, unigrams may not be enough for accurate classification. As an example, the sentence *"I'd never regretted purchasing it"* is actually a positive comment. However, the unigram approach treats *"never"* and *"regretted"* as separate negative words. From our previous experiments, compared with the *unigram* approach, the use of "*unigram with negation phrases"* (e.g., "*never regret*") through simple linguistic processing improved classification accuracy slightly from 76% to 79%.

Other researchers have worked on genre classification. Finn, Kushmerick and Smyth (2002) investigated *automatic genre detection,* which decides whether a document presents the opinion of its author or reports facts (i.e., genre of subjectivity). C4.5, a decision tree induction program (Quinlan, 1993), was used with various text features: *bag of words* (unigrams), *part-of-speech*, and hand-crafted shallow linguistic features (e.g., average sentence length). For the *part-of-speech* approach, a document is represented as a vector of 36 *part-of-speech* features, expressed as percentages of the total number of words for the document. They argued that the *part-of-speech* approach provided the best accuracy when the learned classifiers were generalized from the training corpus to a new domain corpus. Genre detection will be useful for distinguishing review (opinion) documents from different kinds of Web documents.

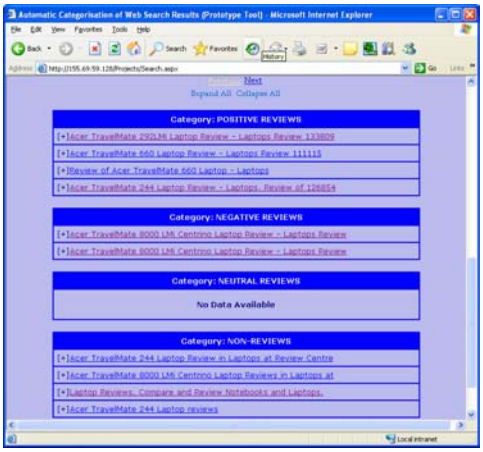
Sentiment-Based Search

We have developed a prototype system that performs sentiment categorization of Web search results. Figure 1 shows an example of a sentiment-based search screen and its result pages. The prototype search interface (Figure 1(a)) allows the user to specify query terms, such as a product name, and passes them to a search engine selected by the user. Search results (Figure 1(b)) are grouped into four categories, namely *Positive*, *Negative*, *Neutral,* and *Non-Reviews*. A positive review (Figure 1(c)) means that the overall sentiment expressed by the document content is positive. *Neutral* means that the system was unable to determine the overall sentiment value for the document, that is, the review

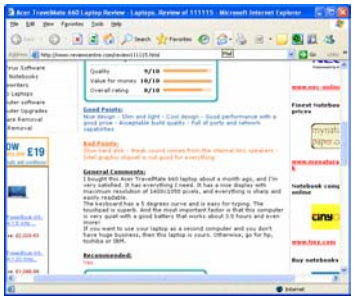
document is neither *positive* nor *negative*. Documents categorized under *Non-Reviews* are considered not relevant as they do not contain any review content. A page deemed by the prototype to be a *non-review* will not be processed for sentiment classification.



(a) User Query



(b) Query Results



(c) Positive review document

Figure 1. An Example of a Sentiment-Based Search and its Results

As shown in Figure 1(a), the user has the option to append "review" or similar words to his/her search terms as this increases the chance of retrieving review documents. We have provided a *site restriction* option, where two Web sites that are built especially for reviews of products can be selected (www.epinions.com and www.reviewcentre.com). The reason for this provision is to increase the likelihood of getting only review documents, because a Web search engine usually returns all sorts of pages. However, the user may also opt for "No restriction".

The function of the *review filtering* option is to 'filter out' non-review documents —to determine if a page returns non-reviews or a list of reviews, instead of the review itself. This prototype system classifies a document as positive or otherwise by analyzing either the summary (i.e., the document snippets from the search engine: the URL, link description, link title, etc.), or the full document content pointed to by the URL—the user has an option to select either. Naturally, the full document option provides better accuracy as it analyzes all the available information from the page. However, this takes a longer time as the system has to fetch the individual documents.

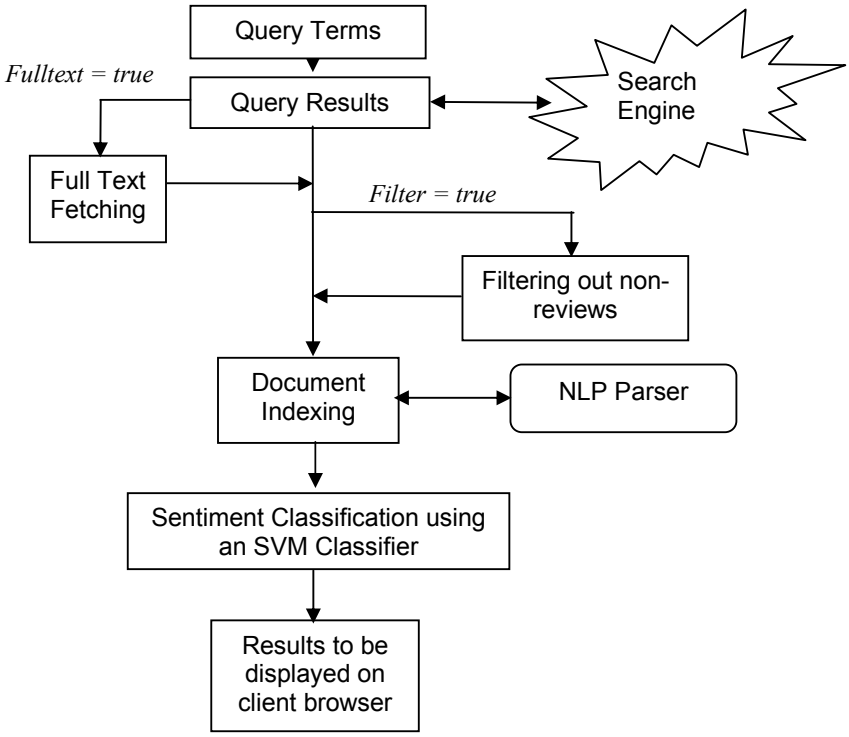


Figure. 2. An Overall Process Flow in the Prototype System

Figure 2 presents an overall process flow in the prototype system. When a user clicks on the *Search* button after he/she inputs a query and selects options, this process flow involves some or all of the following processes in order:

1) Passing the search query to the selected search engine *(Query Terms)*
2) Retrieving the documents from the selected search engine *(Query Results)*
3) Processing the content, and storing them in a database
4) Collecting, cleaning, and storing the full text pointed to by each of the URLs stored in the database if the full text option is chosen *(Full Text Fetching)*
5) Filtering out non-reviews if the review filtering option is set to yes *(Filtering out non-reviews)*
6) Passing either the summary information or the full text to a Natural Language Processing (NLP) parser for word stemming and indexing *(Document Indexing)*
7) Generating input data (document vectors) with stemmed words for classification by an SVM classifier, which was built in our previous study
8) Passing the indexed documents to the SVM classifier for sentiment classification *(Sentiment Classification using an SVM classifier)*
9) Displaying the classified results back to the client browser *(Results to be displayed on client browser).*

Review Filtering

The goal of the review filtering process is to remove pages that may not be review pages. Our prototype system looks for *Regular Expressions* (See Tables 1 and 2) that usually define a non-review page. Once a document is found to be a non-review page, it will not be processed further, other than being displayed in the search results. This means that we will not put this document through the document indexing process or the classification module.

We define a review document as a page that generally contains only a single review, since the sentiment classifier is designed to classify one review at a time. This means that a page that contains more than one review or a list of reviews is filtered out as a non-review. If a page has only the first few lines of the review and requires the user to click on a link to read the rest of the review, this page will also be considered a non-review page. This definition holds true if we make use of the full text for filtering.

The main difficulty in the current prototype system is low sentiment classification accuracy when it takes document snippets instead of whole documents as input, with the no site restriction option. Deciding whether a snippet is a review or non-review is no an easy task, since the snippet information usually does not provide much information with respect to sentiment. Initially, we used a simple algorithm which checks whether some representative words, such as review, occur in the snippets. If it is, it is considered a review snippet. However, this simplistic approach did not work well.

Some researchers have attempted to develop methods to identify the semantic orientation of adjectives using corpus statistics. Hatzivassiloglou and Wiebe (2000) worked on automatic learning of subjective adjectives from corpora, and applied them to distinguish between sentences that present opinions from sentences that objectively present factual information. We use a collection of subjective adjectives (gradable, polar, and dynamic adjectives) from them to filter out non-review snippets. This *subjectivity tagging* can improve our current review filtering function, so that the sentiment classifier performs classification on only the snippets containing opinions (i.e., subjective adjectives).

Generally, the way review filtering is implemented depends heavily on whether we utilize site restriction. Most review sites have specific URL structures for single review pages. By making use of this characteristic, review filtering for such sites is 100 percent accurate. For filtering without site restriction, either the snippet information or the full text is scanned, in two phases. In the first phase, documents that contain the *Regular Expressions* in Table 1 or Table 2 are filtered out as non-reviews. The remaining documents go through phase two to be scanned against a collection of subjective adjectives, and are filtered as reviews if they contain any of the adjectives. The implementation details of review filtering are discussed in the following subsections.

Filtering with Site Restriction

Filtering non-reviews from specific review sites requires analyzing the structure of the URL. For www.epinions.com, if the URL contains either the string "-review- " or "/content", then the URL is that of a single review page, provided that the URL does not also contain the string "/show_~allcom". For www.reviewcentre.com, if the URL contains the string "/review" and not "/reviews", then it is that of a single review page. Thus we need only the snippet information to do filtering. Besides achieving accurate filtering, another advantage of using specific sites is that we know exactly where the review information is. Hence, we are able to extract that part of the document and process only this part for indexing and classification. This would lead to a cleaner input into the classifier module and more accurate classification. This applies to classification based on full text.

Filtering with No Site Restriction

Without site restriction, we can scan either the snippet information or the full text to determine if the document is a review page. Filtering is done in two phases, as explained in the sub-sections that follow.

Using Full Text

In the first phase of filtering, we assume that a page is a review page unless found otherwise. We scan through the document's original text (no HTML tags removed) to look for hyperlinks like "read full review". The existence of such links indicates that the page does not contain the full review but most probably the first few lines of the full review that is on another Web page. Other *Regular Expressions* that make a Web page a non-review are listed in Table 1. Also, a full text description that contains less than 1,000 words is deemed to be a non-review. After the first phase, we will have two sets of documents—non-reviews and reviews. Only reviews will go through the second phase, where the full text is scanned against a collection of subjective adjectives. The idea is that if a textual document or even a simple sentence contains a subjective adjective (e.g. *good*, *bad*), this indicates that a sentiment review text has been found. Documents that go through this phase and are found not to contain any of these adjectives will be regarded as non-reviews.

Using Document Snippets

As long as the snippet contains none of the *Regular Expressions* in Table 2, but contains a sentiment word (a subjective adjective), the prototype system will treat the document as a review.

Table 1. Indication of non-reviews for full text

Regular Expression	Found in	Meaning of Regular Expression
Archive	Link title	Indicates that the page is an archive of reviews
Review form	Link title	Indicates that the page is a blank review form
detailed review.*	Unclean full text	
click here for full review.*	Unclean full text	
read the rest of the review.*	Unclean full text	
read.*full.*review.*	Unclean full text	
Full review.*	Unclean full text	
Has not been reviewed	Full text description	Indicates that no reviews have been submitted for the product
no reviews yet	Full text description	
no review.*submitted	Full text description	
no reviews available	Full text description	
be the first to write a review	Full text description	
subject:	Full text description	Checks for an actual forum page

Table 2. Indication of non-reviews for document snippets

Regular Expression	Found in	Meaning of Regular Expression
Review form	Link title	Indicates that the page is a blank review form
Has not been reviewed	Link description	Indicates that no reviews have been submitted for the product
Has not yet been reviewed	Link description	
No review.*submitted	Link description	
No reviews	Link description	
Be the first to write a review	Link description	

User Evaluation

A preliminary user evaluation was carried out using a questionnaire, soliciting users' assessment and answers to some open-ended questions. Evaluators were asked to perform some product review search tasks using the prototype system. They were also asked to perform a search using their own search scenarios. The evaluation focused on system usability: learnability, flexibility, robustness, and accuracy of the system in terms of document classification according to the overall sentiment. Positive user feedback (with average scores above 4, on a scale of 1 to 5) was received in an informal evaluation of our prototype system, involving eight persons with more than five years of experience in using the Internet.

Table 3. Evaluation results of the prototype sentiment-based meta search engine

ID	Input Combination				Average Accuracy		
	Site Restriction		Information Available		Review Filtering	Sentiment Classification of top 20 items retrieved	Sentiment Classification only of reviews filtered
	epinions.com	None	Document snippet	Full Text			
1	Y			Y	99%	87%	87%
2		Y		Y	76.5%	58%	81%
3		Y	Y		65%	42%	43%

We also conducted a small experiment to evaluate the accuracy of the review filtering and sentiment classification. In the experiment, 10 product queries were submitted to the meta search engine, and the first 20 items retrieved for each query were used to calculate the accuracy of review filtering and sentiment classification. Three of these products were "canon digital ixus 400", "nokia 3310", and "ipod mini".

In Table 3, the first row (ID1) shows the results for a specific Web site (*epinions.com*), using full text from the site. Since the sentiment classifier was trained with review documents from the Review Centre Web site, it was interesting to see how well it fared with documents from another site. The scores were quite high for both review filtering and sentiment classification. For "Sentiment Classification only of reviews filtered", non-review documents not successfully filtered out were not counted in the accuracy calculation (it assumes that review filtering is perfect). The second row (ID2) gives the results for the experiment with no site restriction, using full text. The scores declined somewhat, but still looks promising if we can improve the review filtering function. The third row (ID3) is for an experiment with no site restriction, using document snippets. The sentiment classification accuracy was relatively low (42%), which is only 9% higher than for random guessing (33% each for positive, negative, and neutral). Possible approaches for improving the accuracy include using another classifier trained with a document snippets data set (instead of a full text data set), and also making the review filtering function retain only snippets containing several subjective adjectives (i.e., set a more stringent criterion for review documents). In general, the prototype meta search engine best met its objectives when the user chose a specific site and used full text from the site for processing.

Future Work and Conclusion

Sentiment-based searching is a challenging area that needs further research to meet the needs of users. Using the current prototype system, we plan to explore various aspects and applications of sentiment-based searching on the World Wide Web. The following are some features that can be improved or added in future work:

- More sophisticated review filtering. We will build another classifier to classify a document as a review or non-review. This will be a genre classification of Web documents.
- Improvement in sentiment classification accuracy, especially for the option with no site restriction using summary information.
- Extension to non-electronic products. Most of the user evaluators suggested that it would be nice if the prototype could also cater to sentiment classification of non-electronic products, such as *movies*, *novels*, and *stage plays*.

References

Chen, H., & Dumais, S.T. (2000). Bringing Order to the Web: Automatically Categorizing Search Results, In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'00)* (pp. 145-152).

Finn, A., Kushmerick, N., & Smyth, B. (2002). Genre classification and domain transfer for information filtering. In *Proceedings of ECIR-02, 24th European Colloquium on Information Retrieval Research* (pp. 353-362).

Hatzivassiloglou, V., & Wiebe, J.M. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of 17th International Conference on Computational Linguistics,* Saarbrücken, Germany (pp. 299-305).

Na, J.-C., Khoo, C., Chan, S., and Hamzah, N.B. (2005). A sentiment-based search in digital libraries, In *Proceedings of Joint Conference on Digital Libraries '2005, Denver* (pp. 143-144).

Na, J.-C., Sui, H., Khoo, C., Chan, S., & Zhou, Y. (2004). Effectiveness of Simple Linguistic Processing in Automatic Sentiment Classification of Product Reviews, In *Proceedings of the 8th International Society for Knowledge Organization Conference '2004, London* (pp. 49-54).

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine-learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Philadelphia* (pp. 79-86).

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufman.

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

Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., & Ma, J. (2004). Learning to Cluster Web Search Results, In *Proceedings of the 27th Annual International ACM SIGIR Conference, Sheffield, UK* (pp. 210-217).