2011 Second International Conference on Emerging Applications of Information Technology

Using Negative Information in Search

Sauparna Palchowdhury
CVPR Unit
Indian Statistical Institute
203 B T Road Kolkata 700108, India
sauparna.palchowdhury@gmail.com

Sukomal Pal
Indian Institute of Information Technology
Deoghat, Jhalwa
Allahabad 211012, India
sukomalpal@iiita.ac.in

Mandar Mitra
CVPR Unit
Indian Statistical Institute
203 B T Road Kolkata 700108, India
mandar@isical.ac.in

***Abstract*—Consider a user searching for information on the World Wide Web. If the information need of the user is somewhat specific, and if the user is permitted to provide a detailed description of his precise need, then it is quite likely that this description will include negative constraints, i.e., specifications of what the user is *not* looking for. A search engine that makes use of such constraints is likely to return more accurate results. In this paper, we consider the problem of identifying such negative constraints from verbose queries. A maximum-entropy classifier is trained to identify negative sentences in verbose queries with about 90% accuracy. We next study how retrieval effectiveness is affected when these negative sentences are eliminated from the queries. We find that this step results in modest improvements in retrieval accuracy, but our analysis suggests that significant improvements can be obtained if negative sentences are properly handled during query processing.**

***Index Terms*—search; retrieval; negative constraints; maximum entropy classifier; INEX;**

I. Introduction

Consider a user searching for information on the World Wide Web. If the information need of the user is somewhat specific, a detailed, natural-language description of this need will quite often contain negative constraints. For example, a user may be looking for information on literary works dealing with the partition of India. A precise, natural-language specification of the user's need may read as follows: "I am looking for information about literary works (novels, stories, poetry) that have the partition of India as their subject. *Works set in that period, but not having the partition as their central theme, are not of interest. Also irrelevant are historical / non-fiction accounts about the partition.*"

Most Information Retrieval (IR) systems, however, are unable to handle such detailed queries. Instead, a terse set of keywords is expected as input. Even when a verbose, natural-language query is provided as input, it is usually converted into a set of keywords and keyphrases that is used to retrieve matching documents. In the process, some fine-grained information about the user's need is lost. More specifically, the system does not avoid topics that the user does not want information about. Thus, even when a user precisely describes his/her need using a verbose query, the accuracy of search systems typically does not improve.

Some search engines do provide an "Advanced Search" interface where it is possible to specify keywords that the user does not want. While such a facility is adequate for handling relatively simple queries of the form *I want to know about X but not about Y*, it is not enough for more subtle / complex queries (such as the example given above).

Our broad objective is to explore ways to take into account the detailed specification provided in a verbose query, and provide more accurate search results to the user. In order to effectively and automatically make use of such detailed specifications, we need to address two problems: (i) identifying the constructs (sentences / clauses / phrases) in the detailed information need statement that specify negative constraints; and (ii) utilising this information in the retrieval process. In this paper, we describe our attempts to identify negative information in a verbose query, and show how retrieval results are affected when these negative portions are simply removed from the user's query. We first use a supervised Machine Learning technique to identify negative sentences from verbose queries in a query collection and purge these sentences from the queries (Section IV). The modified queries are then used to retrieve documents from a standard benchmark corpus. We find that using the modified queries results in some improvements in retrieval accuracy. We also analyze these results in greater detail and identify some weaknesses in our approach (Section V). A significant boost in retrieval performance should be achievable if these weaknesses are addressed.

II. Related Work

Presence of negation in text has been made use of in detecting contradiction, sentiment analysis and in finding absence of medical conditioins in medical reports. In the latter case, several methods have been used, like, ad hoc classification [1], regular expression matching [2], lexical analysis and parsing [3], and Machine Learning classifiers. Goryachev et al. provide a survey of these efforts [4]. In our work, we use an off-the-shelf Maximum-Entropy classifier (see Section IV for details). Another use of this classifier is reported in [5].

Previous work on verbose queries have applied various query-shortening techniques like key-concept extraction, improving ranking by weighting terms and query quality predictors. The importance of verbose queries is exemplified by the results of an analysis done on a month's log from MSN search [6] which showed that 10% of the 15 million queries are longer than 5 words. A recent report by Huston and Croft [7] compares all these verbose-query techniques we mentioned.

978-0-7695-4329-1/11 $26.00 © 2011 IEEE
DOI 10.1109/EAIT.2011.86

III. Benchmark Collection

We use benchmark datasets provided by the Initiative for the Evaluation of XML Retrieval (INEX) [8], [9]. These datasets consist of a collection of documents (in which we search for information), sample queries, and relevance judgments (information about which documents are relevant for which queries).

a) Documents: The INEX 2006 corpus consists of 659,388 full-text articles crawled from the English Wikipedia[1] and is about 4.6 GB in size. Images are removed, and the text is marked up using XML. The INEX 2009 collection contains more than 2.6 million documents and is about 50.7 GB in size.

b) Queries: Each query contains the information need expressed in natural-language, formulated by the INEX participants themselves. The contents are marked up by XML consisting of a *title*, a *description* and a *narrative* field. These fields provide progressively more detailed specifications of the user's information need. The title and the description are brief, but the narrative is usually verbose and may contain negative constraints. An example query is given below.

```
<topic id="2009080" ct_no="268">

<title>international game show formats</title>

<description>I want to know about all the game show formats
that have adaptations in different countries.</description>

<narrative>Any content describing game show formats with
international adaptations are relevant. National game shows
and articles about the players and producers are not
interesting.</narrative>

</topic>
```

Fig. 1. Query 80 from the INEX 2009 ad hoc query collection.

IV. Detection and Separation of Negative Information

A. Defining Positive and Negative Sentences

Sentences specifying what the user wants are termed as *positive*, while those which specify what the user is not looking for are termed *negative*. Thus, the sentence *I need X but I don't want Y*, when spliced, gives a positive part *I need X* and a negative part *don't want Y*. A sentence of the form *Z is not irrelevant* is labeled as positive because of the presence of double negation. A query broken into its positive and negative parts is shown in Table I.

a) Manual Separation: We used the INEX ad hoc query collections of years 2007, 2008 and 2009, containing respectively 130, 135, and 115 queries, for our experiments. The narrative of each query was broken into sentences and manually labeled as positive and negative.

[1]http://en.wikipedia.org

TABLE I
Positive and negative parts of Query 80.

Whole query
Any content describing game show formats with international adaptations are relevant. National game shows and articles about the players and producers are not interesting.
Positive part
Any content describing game show formats with international adaptations are relevant.
Negative part
National game shows and articles about the players and producers are not interesting.

b) Machine Classification: We used the Stanford Classifier [10], a Java implementation of a Maximum-Entropy Classifier. The features used were n-grams of length 1 to 4 and prefix-suffix n-grams. The manually labeled set of sentences (see *a)* above) was used to train the classifier, and then classify a test-set of sentences and phrases into one of two classes.

B. Results

Table II shows the results obtained using the above classifier. We tested the classifier separately on the INEX 2008 and INEX 2009 query collections. The training set, overall accuracy, and percentage of mis-classifications are shown in the table for each of these collections. We find that the classifier performed fairly well, achieving approximately 90% accuracy.

TABLE II
Classifier performance. + *to* - indicates positive sentences wrongly classified as negative (and vice-versa)

Test set	Accuracy	- to +	+ to -	Training set
2008	90.3%	6.8%	3.0%	2007
2009	91.5%	5.4%	3.1%	2007, 2008

Negative sentences (or phrases) having an uncommon pattern were wrongly classified. Sentences containing double negations also had a propensity to fall into the negative class. Table III shows sentences with rarely occurring, complicated, negative phrases like *'not useful'*, *'only if'*, *'far off'*, *'not interested'*, *'without'*, *'area of interest excludes'*, *'not relevant unless'*. The first two sentences in table IV have a negation that qualifies something that is not the key idea expressed in the query. The last two have a double negation and hence a positive meaning.

TABLE III
Negative; wrongly classified as positive

"The area of interest excludes information related to intrusion prevention systems or any other reactive-based approaches"

"But an element that describes post ww2 imperialism is far off"

"I'm not interested in wine-based drinks, profession related to wine and so one"

"However, components about culture, economy, geography or demographics of Tibet are not relevant, unless they highlight some features of the independence movement"

TABLE IV
POSITIVE; WRONGLY CLASSIFIED AS NEGATIVE

"My goal is not to write a report, i am just want to read about the war so all information is relevant : battles descriptions, organizations, dates, events, historian people and parties etc"

"However I am not an engineer, therefore I would like to know the basic concepts of different kind of electric hybrid vehicles, particularly battery-powered ones"

"... or the information which is not unique to Japanese food are regard as non-relevant"

"I'm not interested in different breweries and brands of ale if they don't explain about different types of ale and what makes them different"

V. RETRIEVAL AND EVALUATION

A. Query Sets

Since our main aim is to study the impact of removing negative sentences from verbose queries, we first eliminate from the INEX 2008 and INEX 2009 query collections all queries that do not contain any negative sentences.

Corresponding to a given query q, we now construct a positive query q_p by combining the title, the description, and all positive sentences from the narrative of q. A negative query q_n is similarly constructed by including only negative sentences from the narrative of q. Three sets of queries were formed from each query collection for our experiments: the unmodified queries (denoted Q), the set of positive queries (P), and the set of negative queries (N). P and N correspond to the manually labeled queries; their machine-labeled counterparts, obtained using the classifier, are denoted by P_M and N_M.

B. Retrieval

We use the SMART [11] IR system, with the Lnu.ltn term-weighting scheme (see [12] for details) for indexing and retrieval. As a baseline, 1500 documents are retrieved for each query (this is as per the specification of the INEX ad hoc task). We also use blind feedback to obtain another set of retrieval results [12]. Results are quantitatively evaluated in terms of MAP (Mean Average Precision, a standard metric used to measure retrieval accuracy [13]).

1) Overall Results: Retrieval results obtained using Q, P, N, and P_M are shown in Table V. The table shows absolute MAP values as well as the percentage changes between the Q and P results, and the Q and N results. The baseline results are denoted by 'b', while the rows labeled 'fb' show the performance obtained when blind feedback is also used.

We expected P (and P_M) to yield better results than Q, and N (also N_M) to result in poorer performance compared to Q. The premise of this assumption was that the absence of negative sentences (and therefore keywords related to unwanted topics) in P would bring up a larger number of relevant documents compared to Q. For analogous reasons, N is expected to bring up a larger number of irrelevant documents.

2) Per-Query Results: In this section, we analyze the results obtained at the individual query level for the INEX 2008 query collection. The INEX 2009 query collection has similar properties, but the details for this collection are not shown for lack of space. The majority of the queries in P show an improvement, whereas the majority in N perform poorly. P_M performs almost as well as P. Table VI tabulates the number of queries for which performance deteriorates / improves when P, N and P_M are used in place of the original Q. Figures 2 and 3 display the same information graphically.

TABLE V
OVERALL MAP. FIGURES IN () SHOW % CHANGE W.R.T. Q.

INEX year	run	Q	P	N
2008	b	0.2586	0.2660 (2.9%)	0.2265 (-1.2%)
(44 queries)	fb	0.2706	0.2827 (4.5%)	0.2496 (-7.8%)
2009	b	0.2499	0.2642 (5.7%)	0.2348 (-6.0%)
(36 queries)	fb	0.2504	0.2651 (4.4%)	0.2382 (-4.9%)
INEX year	run	Q	P_M	N_M
2008	b	0.2564	0.2624 (2.3%)	0.2397 (-6.5%)
(31 queries)	fb	0.2638	0.2748 (4.2%)	0.2574 (-2.4%)
2009	b	0.2728	0.2790 (2.3%)	0.2768 (1.5%)
(36 queries)	fb	0.2814	0.2897 (2.9%)	0.2914 (3.6%)

TABLE VI
NUMBER OF INEX 2008 QUERIES SHOWING A CHANGE IN PERFORMANCE.

Set	Deteriorates	Improves
P	11	33
N	34	10
P_M	8	23

We next analyze the 'anomalous' queries, i.e., queries in P that degrade and those in N that improve performance. The fall of performance of 555, 582 and 626 in P, is attributed to the loss of terms when pruning the query from Q to P. For example, in query 555, the user asks for *photos of Amsterdam*. The query text goes on to describe unwanted topics related to Amsterdam, using this term in phrases that belong to negative sentences, like *'American cities with the same name as **Amsterdam**'*, *'ships called **Amsterdam**'*, and so on. The positive query loses these phrases to the negative counterpart and consequently, the term weight of *'Amsterdam'* decreases. Similarly, for query 626, the term *'classifier'* is absent from the positive query, but the original query seeks information on *applications of Bayes filters* of which a Bayesian **classifier** is an example. So documents on Bayesian classifiers, though relevant to the original query, are ranked lower when the positive query is used.

To explain the improvement of 609, 610 and 669 in N, we observe that query pruning — even pruning of positive sentences — sometimes has the effect of improving performance by eliminating noisy terms from the verbose versions. The query vectors of the negative queries have less than half the terms contained in the original, but retain important terms. Thus, the query focus actually improves in these cases, and relevant documents are promoted to better ranks in the retrieval results.

In Figure 4 we charted the 27 queries in $|P \cap P_M|$. The queries in P_M do well to achieve the performance of those in

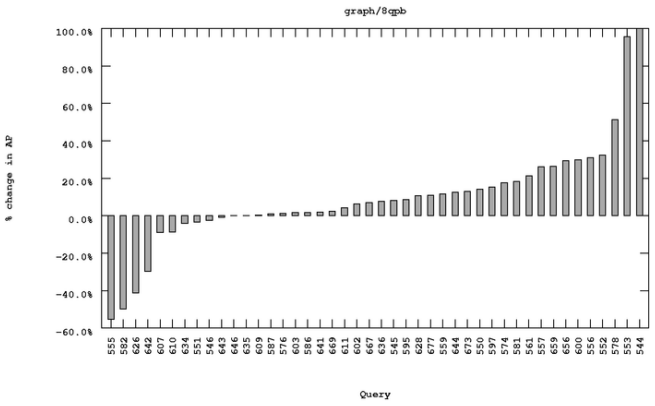

Fig. 2. Performance of each query in set P. % change in Average Precision (AP) is plotted for the 44 queries. The change is computed with respect to their counterparts in Q.

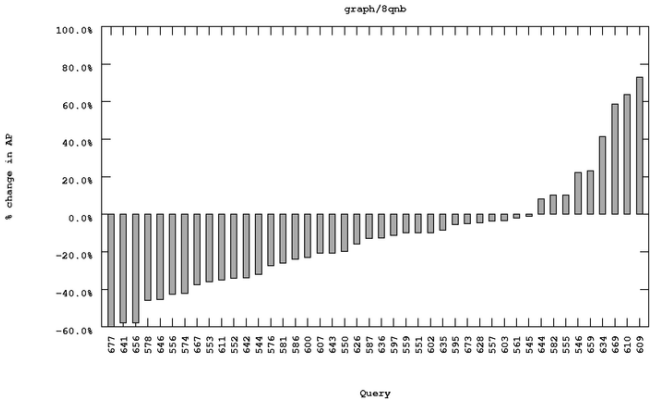

Fig. 3. Performance of each query in set N. % change in AP is plotted for the 44 queries. The change is computed with respect to their counterparts in Q.

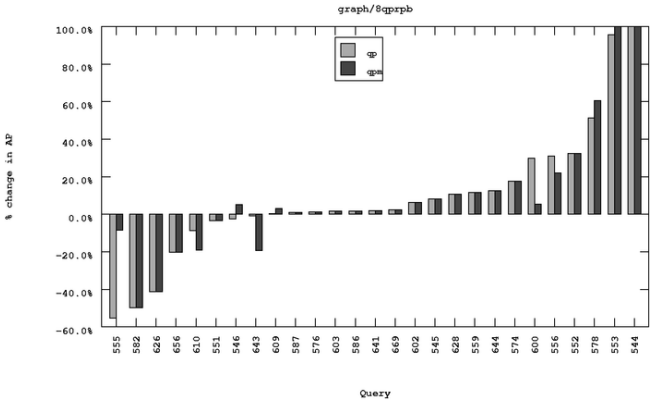

Fig. 4. Comparison of the performance of P_M with P.

P, which can be attributed to the classifier's performance. At more than 90% accuracy, it created set P_M with most of its queries identical to those in Q.

VI. Limitations and Future Work

When addressing the first of our sub-problems, we restricted ourselves to using *sentences* as units of negative information. Sub-sentence level negation detection is not being done. As a result, key terms may be lost as we saw in Section V-B2. Our second sub-problem has been addressed in a fairly simple way. The IR system treats a keyword in isolation ignoring the words surrounding it in the query. These words are instrumental in defining the keyword's negative or positive sense. We need a way to convey the semantics of natural-language to the IR system. Also, the retrieval granularity could make a difference, i.e., retrieving passages instead of whole documents can allow us to match the positive queries to relevant portions within a document, without penalising the document as a whole because of the coexistence of negative terms.

VII. Conclusion

In this work we have demonstrated a way of approaching our broad objective of making use of the detailed specifications in verbose queries to improve search results. Our focus was on removing negation to improve verbose queries. Using a classifier we were able to automatically detect and remove negation and show that this improves accuracy across the majority of queries. Scope for improvement remains in making use of the semantics of the language, and in using negative information in more sophisticated ways than simply removing negative sentences.

References

[1] D. Aronow, F. Feng, and W. Croft, "Ad hoc classification of radiology reports," *J. AMIA*, vol. 6, no. 5, pp. 393–411, 1999.

[2] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries," *J. Biomed. Inform.*, vol. 34, no. 5, pp. 301–310, 2001.

[3] P. Mutalik, A. Deshpande, and P. Nadkarni, "Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the umls.(negfinder)," *J. AMIA*, vol. 8, no. 6, pp. 598–609, 2001.

[4] S. Goryachev, M. Sordo, Q. T. Zeng, and L. Ngo, "Implementation and evaluation of four different methods of negation detection," Tech. Rep., 2006.

[5] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," http://www.stanford.edu/~alecmgo/papers/TwitterDistantSupervision09.pdf.

[6] M. Bendersky and W. B. Croft, "Analysis of long queries in a large scale search log," in *Proc. 2009 workshop on Web Search Click Data (WSCD 09)*, 2009, pp. 8–14.

[7] S. Huston and W. B. Croft, "Evaluating verbose query processing techniques," in *Proc. 33rd ACM SIGIR*, 2010, pp. 291–298.

[8] S. Geva, J. Kamps, and A. Trotman, Eds., *Advances in Focused Retrieval*, ser. LNCS, vol. 5631, 2009.

[9] ——, *Focused Retrieval and Evaluation*, ser. LNCS, vol. 6203, 2010.

[10] A. Rafferty and C. Manning, "Stanford classifier," http://nlp.stanford.edu/software/classifier.shtml.

[11] G. Salton, Ed., *The SMART Retrieval System—Experiments in Automatic Document Retrieval*. Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

[12] S. Pal, M. Mitra, and D. Ganguly, "Parameter tuning in pivoted normalization for XML retrieval: ISIINEX09 adhoc focused task," in *Focused Retrieval and Evaluation*, ser. LNCS, no. 6203, 2010, pp. 112–121.

[13] C. D. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*. Cambridge University Press, 2009.