Using Negative Information in Search

Sauparna Palchowdhury
Sukomal Pal
Mandar Mitra

Indian Statistical Institute
203 B T Road, Kolkata 700108
West Bengal, India

February 18, 2011

Introduction

Problem

- *Verbose* queries give users more latitude.
- Queries may contain *negation*, i.e. specifications of what is *not* wanted.
- Search engines use keyword matching rather than query understanding $\Rightarrow$ keywords from negative portions are also used for matching.
- Does retrieval effectiveness improve on removing negation ?

A Verbose Query with Negative Information

"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.*"

Related Work

- An MSN search log showed 10% of 15 million web queries to be longer than 5 words.
- Query shortening techniques have been used.
- Identifying negation in medical reports.
- Sentiment analysis involves finding negative connotations.

Benchmark Collection

INEX - Initiative for the Evaluation of XML retrieval.

- *Corpus* - Full-text articles crawled from the Wikipedia.
  - · 2006 corpus : 659,388 documents, 4.6GB.
  - · 2009 corpus : 2.6 million documents, 50.7GB.
- *Queries* - Natural language queries formulated by INEX participants.
  - · 2007, 2008, 2009 query sets total 380 queries.

Sample INEX Query

```
<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>
```

Detection and Separation of Negative Information

Positive and Negative Parts of a Query

Whole query

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.*

Positive part

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.

Negative part

Also irrelevant are historical / non-fiction accounts about the partition.

Separation Using a Classifier

- A Maximum-Entropy Classfier was trained on manually separated query sets.
- Tested on 2008, 2009 sets.

Table: 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

Retrieval and Evaluation

- The SMART retrieval engine.
- Vector space model.
- MAP (Mean Average Precision) is the evaluation metric.

Overall Results

Table: 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%)

Per-Query Results

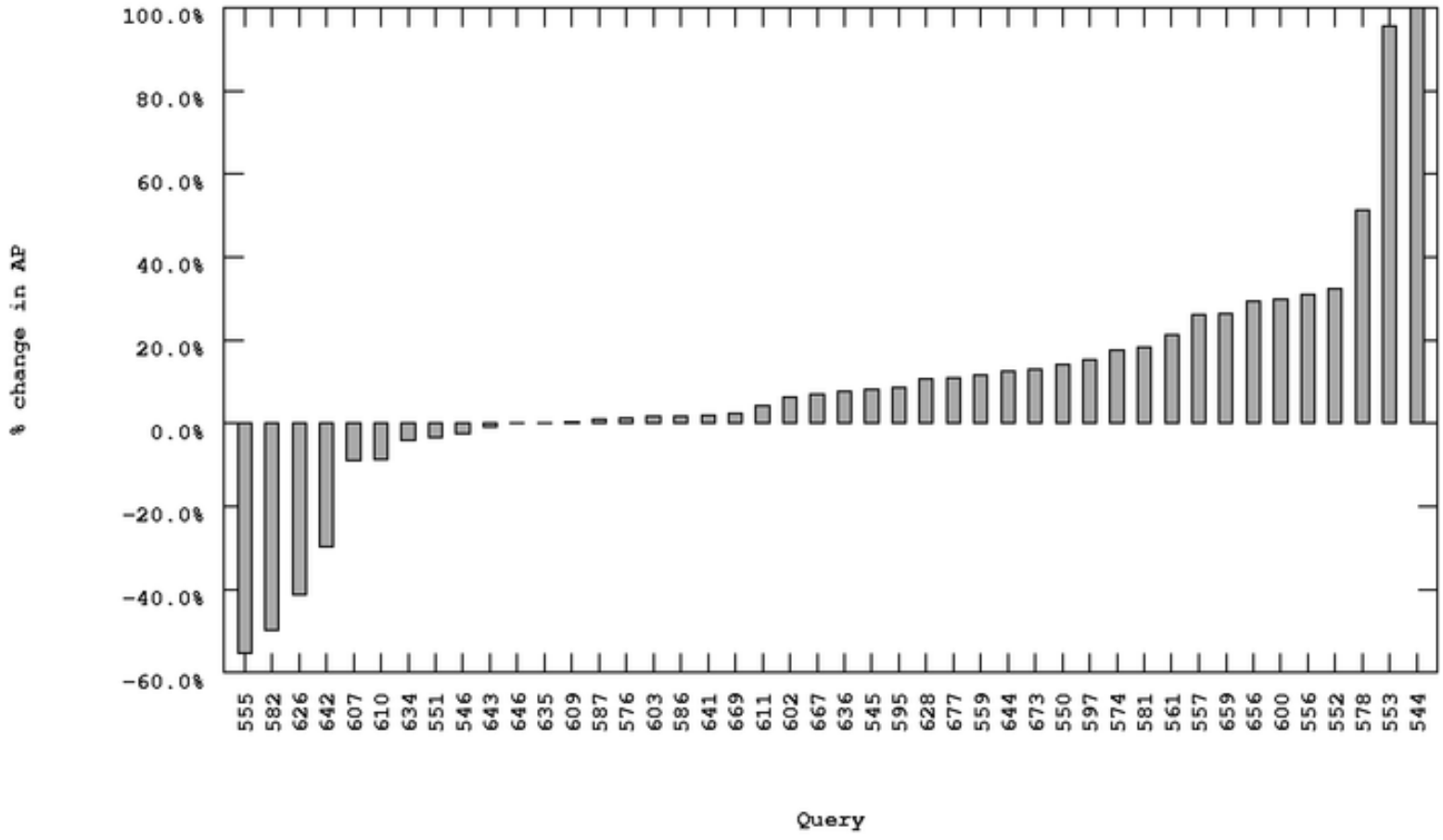

Figure: Performance of each query in set P. % change in Average Precision (AP) is plotted for the 44 queries. The change is computed with respect to their counterparts in Q.

Per-Query Results

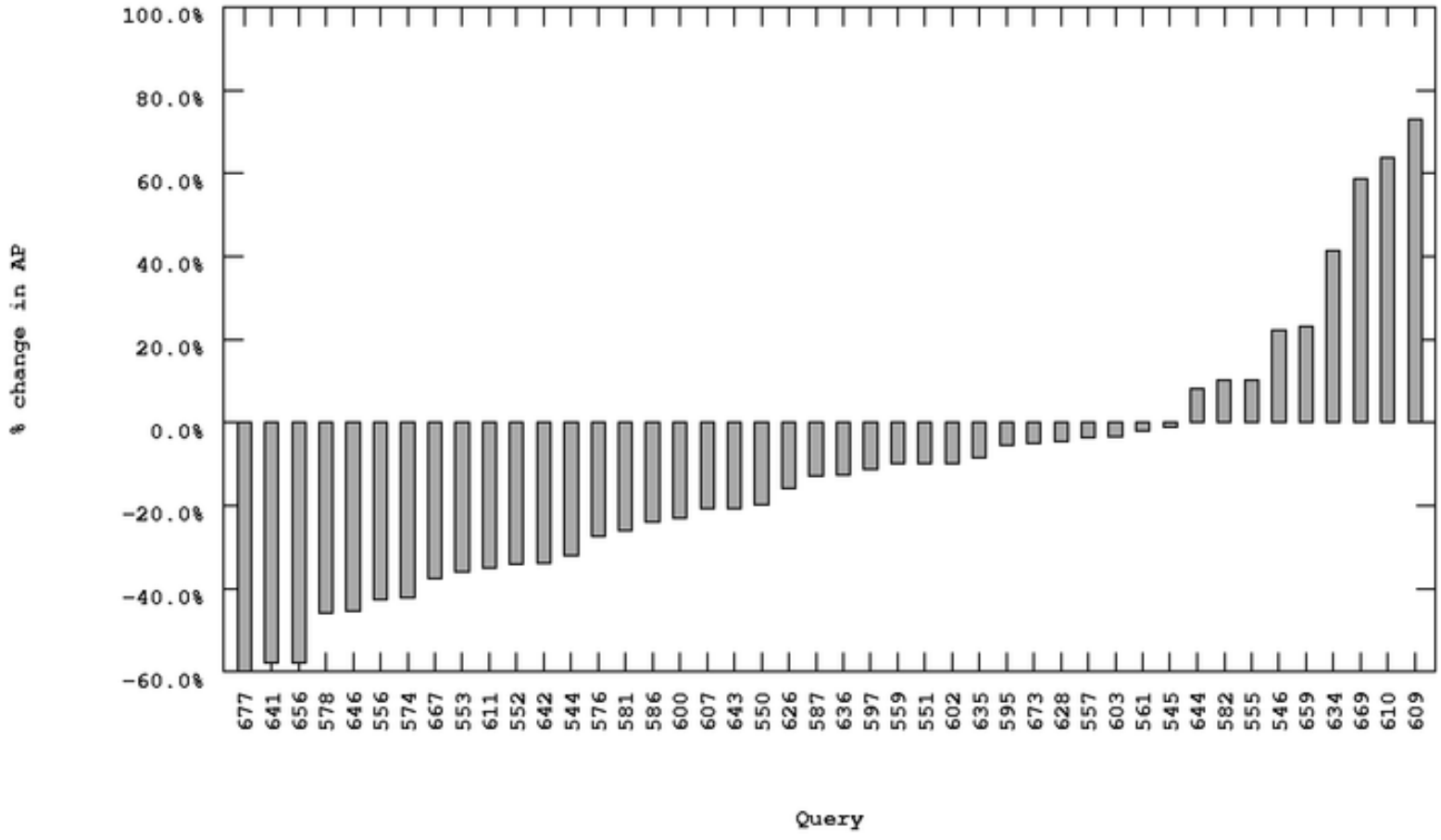

Figure: Performance of each query in set N. % change in AP is plotted for the 44 queries. The change is computed with respect to their counterparts in Q.

Per-Query Results

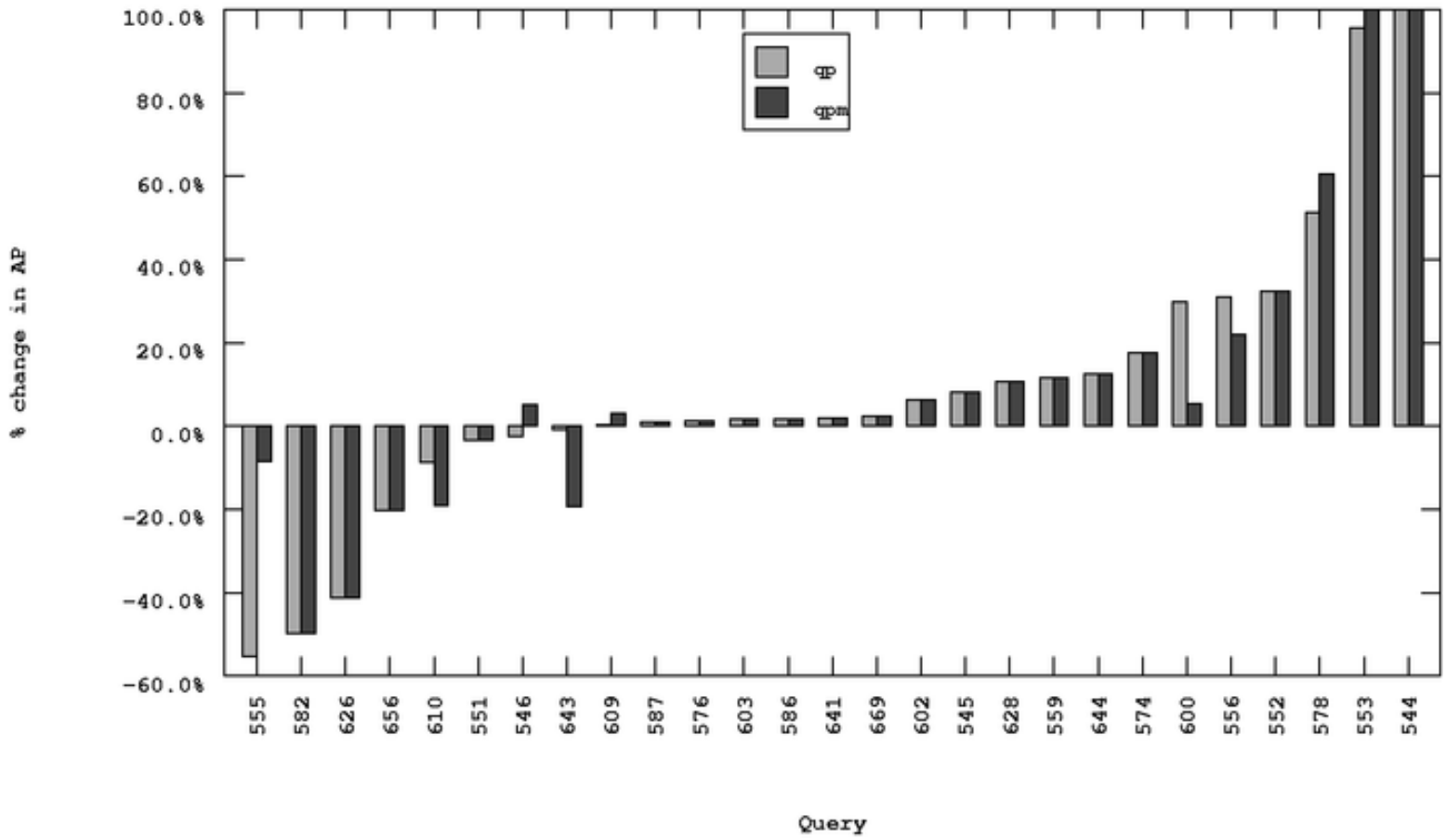

Figure: Comparison of the performance of P_M with P.

Conclusion

Limitations

- Simplistic approach.
- Complicated negative-phrases not dealt with.
- A relatively small number of queries had both a positive and negative part. Larger, more varied sets may have provided further insight.

Future Work

- Affecting term weights.
- Increasing the granularity of the corpus.

Thank you.