Measuring Opinion Relevance in Latent Topic Space

Wei Cheng\textsuperscript{1}    Xiaochuan Ni\textsuperscript{2}    Jian-Tao Sun\textsuperscript{2}
Xiaoming Jin\textsuperscript{3}    Hye-Chung Kum\textsuperscript{1}    Xiang Zhang\textsuperscript{4}
Wei Wang\textsuperscript{1}

\textsuperscript{1}Department of Computer Science, University of North Carolina at Chapel Hill, USA
\textsuperscript{2}Microsoft Research Asia, Beijing, China
\textsuperscript{3}School of Software, Tsinghua University, Beijing, China
\textsuperscript{4}Department of Electrical Engineering and Computer Science, Case Western Reserve University, USA

Advisor: Prof. Wei Wang
IEEE International Conference on Social Computing, MIT, 2011
What is opinion retrieval?

Given a query, find documents that have subjective opinions about the query

Example

A query “ipad”
Relevant: “iPad is perfect.”
Irrelevant: “The lowest price of iPad is $499.”
Routine Solution for Opinion Retrieval: Two-stage Methodology

Corpus

Traditional IR Engine

Initial topically relevant documents

Opinion Analysis

ScoreIR

query

ScoreIop

Opinion relevant documents
Related Work

Opinion identification

- Lexicon-based methods ([1],[2],[3],[4]), i.e., “perfect”, “wonderful”, etc.
- Classification-based methods ([5],[6]), i.e., SVM classifier
- Language model based methods ([7]), i.e., “I...wish”, “I...felt”, “I...enjoyed”, etc.

Identifying opinion targets

- Words distance-based approximate methodology ([3])
- Sentences distance-based approximate methodology ([8])

Integrating topic relevance and opinion for ranking

- linear combination ([8])
- quadratic combination ([3])
Motivation

In the literature, the opinion relevance is handled by very straightforward methods:

1. In [3], the authors count the number of opinion-carrying words (like “perfect”, “terrible”, etc.) around search query terms;
2. W. Zhang et al. [8] count opinion-carrying sentences (identified by sentiment classifier) around the search query.

The drawbacks of this kind of solutions:

- Opinions around the search query may not be related to the query.
- The terms in search query cannot represent all the semantics of the search topic.
- To sum up, topic and opinion are mismatched.
Drawbacks

Example

1 I got a second job at Ritz Camera Shop. 2 I'll be getting many more hours, and making seven dollars an hour plus commission on anything I sell, so probably about 10 dollars an hour. 3 It'll be quite nice working with stuff that I'm into. 4 My brother has pink eye and he's, no doubt, been probing that thing all day whilst sitting here, typing on these very keys. 5 Oh, Casey and I went to see March of the Penguins last night. 6 Penguins are remarkable. 7 I kept making a lot of little noises all throughout the film because it was pretty much killing me. 8 A penguin documentary? 9 I'm so happy for you that you got that job! 10 It'll be so good for you since you are into photography. 11 I wish I could get a job at a gallery or something. 12 That would be helpful down the line. 13 Congratulations on your new job! 14 I wanted to see March of the Penguins...[15 I heard it was super good, the documentary seems wonderful.] 16 The film was dubbed by Morgan Freeman. 17 I was just dropping by and saying hi. 18 I haven't seen you in forever.

Figure: Example document from the TREC Blog06 corpus with Query “March of the penguins”
Relationship among different document sets

Figure: Relationship among $SET_{opinion\_to\_Q}$, $SET_{relevant\_to\_Q}$ and $SET_{opinion}$
Key challenge of opinion retrieval task

- \( SET_{opinion\text{-}to\text{-}Q} \) is a proper subset of \( SET_{relevant\text{-}to\text{-}Q} \cap SET_{opinion} \).
- \( SET_{relevant\text{-}to\text{-}Q} \) and \( SET_{opinion} \) can be retrieved by using traditional IR techniques and opinion mining technologies respectively.
- The crucial challenge of opinion retrieval task is in fact to select out \( SET_{opinion\text{-}to\text{-}Q} \) from \( SET_{relevant\text{-}to\text{-}Q} \cap SET_{opinion} \).

Key point: Measuring opinion relevance

Example

```
Opinion Irrelevant
Query: iPad
Document
Paragraph1: iPhone
having opinions about iPhone
Paragraph2: iPad
without opinions about iPad
```

Wei Cheng, Xiaochuan Ni, Jian-Tao Sun, Xiaoming Jin, Hye-Chung Kum, Xiang Zhang, Wei Wang
Measuring Opinion Relevance in Latent Topic Space
Our overall pipeline for topic model based opinion retrieval

- **Corpus**
- **Traditional IR Engine**
- **PLSA (top ranked)**
- **Topic Space**
- **Opinion relevant documents**
- **Opinion bearing sentences**
- **Projection**
- **Re-ranking**
- **Similarity Evaluating**
- **Query**

**Measuring Opinion Relevance in Latent Topic Space**
Getting initial topically relevant documents

**BM-25**

\[
Score_{IR} = \sum_{t \in Q, D} \ln \left( \frac{N - df + 0.5}{df + 0.5} \right) \cdot \frac{(k_1 + 1)tf}{(k_1(1 - b) + b \frac{dl}{avdl}) + tf} \cdot \frac{(k_3 + 1)qtf}{k_3 + qtf}
\]

where \(tf\) is the term’s frequency in document, \(qtf\) is the term’s frequency in query, \(N\) is the total number of documents in the collection, \(df\) is the number of documents that contain the term, \(dl\) is the document length (in bytes), \(avdl\) is the average document length, \(k_1\) (between 1.0-2.0), \(b\) (usually 0.75), and \(k_3\) (between 0-1000) are constants.
After we obtain the working set (i.e., top ranked documents from initially retrieved documents), we will use Probabilistic Latent Semantic Analysis (PLSA) for inferring a topic space.

PLSA [9] models each document as a mixture of topics. It generates documents with the following process:

- Select a document $d$ with probability $p(d)$
- Pick a latent topic $z$ with probability $p(z|d)$
- Generate a word $w$ with probability $p(w|z)$
Project search query and opinion carrying sentences into topic space

Project opinion carrying sentence

\[ p(s_{op}|z) = \langle p(s_{op}|z_1), \ldots, p(s_{op}|z_K) \rangle \]  

where \( p(s_{op}|z_k) \) is the probability for topic \( z_k \) to generate \( s_{op} \) [10].

\[ p(s_{op}|z_k) = \frac{1}{K'} \cdot \sum_{j=1}^{|V|} n(s_{op}, w_j)p(w_j|z_k) \]  

Project search query

\[ p(q|z) = \langle p(q|z_1), \ldots, p(q|z_K) \rangle \]  

where \( p(q|z_k) \) is the probability for topic \( z_k \) to generate query \( q \).

\[ p(q|z_k) = \frac{1}{K''} \cdot \sum_{j=1}^{|V|} n(q, w_j)p(w_j|z_k) \]
Opinion relevance measurement at the topic level

\[ p(q|s_{op}) = p(Q_{\text{topic}}|s_{op}) = \cos\left(\frac{p(q|z) \cdot p(s_{op}|z)}{\|p(q|z)\|_2 \times \|p(s_{op}|z)\|_2}\right) \]  

Note that \( p(Q_{\text{topic}}|s_{op}) \) is not a strict probability distribution, it’s used to denote the score of the similarity between \( s_{op} \) and \( Q_{\text{topic}} \).
Re-ranking

**Opinion score**

\[
Score_{IOP} = \sum_{s_m} p(q|s_m) \cdot f(s_m) \quad s.t. \quad (1) \quad \text{and} \quad (2), \quad \text{with}:
\]

(1) \quad p(q|s_m) > \mu

(2) \quad s_m \in S_o

where \( S_o \) is the opinion expressing sentence set of document \( d \) and \( f(s_m) \) denotes the sentiment strength of opinion expressing sentence \( s_m \)

**Overall score for ranking**

\[
Score = \text{Combination}(Score_{IR}, \widehat{Score_{IOP}})
\]

**Two combination methods in combining \( Score_{IR} \) and \( \widehat{Score_{IOP}} \):**

- linear combination \( \lambda \cdot Score_{IR} + (1 - \lambda) \cdot \widehat{Score_{IOP}} \) [11, 8];
- quadratic combination \( Score_{IR} \cdot \widehat{Score_{IOP}} \) [3].
Evaluation Setup

Data Sets

- We use 150 queries from Blog06, Blog07 [13], and Blog08 [14] for evaluation.
- We will tune algorithm parameters with Blog Track 06 topics while using Blog Track 07 and 08 new topics as the testing data.
MAP (mean average precision), P@10 (precision at top 10 results) and R-prec (R-Precision)

\[
\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad (9)
\]

\(P@10\) is the precision that considers only the topmost 10 results returned.

\[
\text{AveP}(q) = \frac{\sum_{k=1}^{n} (P@k \cdot \text{rel}(k))}{\text{number of relevant documents}} \quad (10)
\]

where \(\text{rel}(k)\) is an indicator function equaling 1 if the item at rank \(k\) is a relevant document, zero otherwise, \(P@k\) is the precision at cut-off \(k\) in the list.

Example

For this query result, the \(\text{AveP}(\text{average precision})\) is \(\frac{1}{3} \left( \frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right)\) because \(P@1=1/1, P@3=2/3, \text{and } P@5=3/5\).
Evaluation Setup

Evaluation metrics

\[ MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \] (11)

*R - Precision* is precision at \( R \)-th position in the ranking of results for a query that has \( R \) relevant documents.
Experiment environment setting

- **Opinion identification policy:** We will follow M. Zhang et al. [3] and use sentiwordNet [4]. Choose opinion-bearing words with positive/negative strength more than 0.6.
- **Combination policy:** linear combination
- **Size of the working set:** $M=1000$

### Example

**SentiwordNet Sample**

<table>
<thead>
<tr>
<th>Positive</th>
<th>Negative</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>happy 0.875</td>
<td>abject 0.75</td>
<td>unfixed 0.125</td>
</tr>
<tr>
<td>fortuitous 0.5</td>
<td>disastrous 0.75</td>
<td>detached 0.125</td>
</tr>
<tr>
<td>good 0.25</td>
<td>dispossessed 0.875</td>
<td>at_liberty 0.125</td>
</tr>
<tr>
<td>providential 0.875</td>
<td>pathetic 1</td>
<td>ascetic 0.625</td>
</tr>
<tr>
<td>lucky 0.875</td>
<td>ill-fated 0.875</td>
<td>deathlike 0.125</td>
</tr>
</tbody>
</table>
Baseline methods and implementations

**bag-of-words-Method**

\[
Score_{IOP} = \sum_{s_m \in S_o} f(s_m)
\]  

**single-sentence-Method**

\[
Score_{IOP} = \sum_{s_m} f(s_m) \quad \text{s.t.} \quad (1) \quad \text{and} \quad (2), \quad \text{with:}
\begin{align*}
(1) & \quad s_m \in S_o \\
(2) & \quad s_m \text{ contains } q
\end{align*}
\]  

**window-Method**

\[
Score_{IOP} = \sum_{s_m} f(s_m) \quad \text{s.t.} \quad (1) \quad \text{and} \quad (2), \quad \text{with:}
\begin{align*}
(1) & \quad s_m \in S_o \\
(2) & \quad \exists s_i \in S_{neighbors\_of\_s_m} \land s_i \text{ contains } q
\end{align*}
\]
Effectiveness of number of topics

Figure: MAP evolution over topic number
Effectiveness of our framework

1 I got a second job at Ritz Camera Shop. 2 I'll be getting many more hours, and making seven dollars an hour plus commission on anything I sell, so probably about 10 dollars an hour. 3 It'll be quite nice working with stuff that I'm into. 4 My brother has pink eye and he's, no doubt, been probing that thing all day whilst sitting here, typing on these very keys. 5 Oh, Casey and I went to see March of the Penguins last night. 6 Penguins are remarkable. 7 I kept making a lot of little noises all throughout the film because it was pretty much killing me. 8 A penguin documentary? 9 I'm so happy for you that you got that job! 10 It'll be so good for you since you are into photography. 11 I wish I could get a job at a gallery or something. 12 That would be helpful down the line. 13 Congratulations on your new job! 14 I wanted to see March of the Penguins... 15 I heard it was super good, the documentary seems wonderful. 16 The film was dubbed by Morgan Freeman. 17 I was just dropping by and saying hi. 18 I haven't seen you in forever.

Figure: Test document $d$

(a) Query

March of the Penguins

(b) Topical relevance of document $d$

Measuring Opinion Relevance in Latent Topic Space

Wei Cheng, Xiaochuan Ni, Jian-Tao Sun, Xiaoming Jin, Hye-Chung Kum, Xiang Zhang, Wei Wang
Comparison to baselines

**Figure:** MAP & P@10 comparison for Blog Track 06, 07 and 08 topics with different combination parameters

Wei Cheng, Xiaochuan Ni, Jian-Tao Sun, Xiaoming Jin, Hye-Chung Kum, Xiang Zhang, Wei Wang

Measuring Opinion Relevance in Latent Topic Space
Figure: MAP improvement after re-ranking for individual 150 topics of Blog Track 06, 07 and 08
## Comparison with state-of-the-art methods

<table>
<thead>
<tr>
<th>Data set</th>
<th>Method</th>
<th>MAP</th>
<th>P@10</th>
<th>R-prec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blog 06</td>
<td>Best title-only-run at Blog06 [12]</td>
<td>0.1885</td>
<td>0.512</td>
<td>0.2771</td>
</tr>
<tr>
<td></td>
<td>Our Relevant Baseline</td>
<td>0.186</td>
<td>0.32</td>
<td>0.2623</td>
</tr>
<tr>
<td></td>
<td>Our framework</td>
<td>0.2504</td>
<td>0.532</td>
<td>0.3169</td>
</tr>
<tr>
<td></td>
<td>M. Zhang et al. improvement [3]</td>
<td>28.38%</td>
<td>44.86%</td>
<td>16.00%</td>
</tr>
<tr>
<td></td>
<td>K. Seki et al. improvement [7]</td>
<td>22.00%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Our framework’s improvement</td>
<td><strong>34.62%</strong></td>
<td><strong>66.25%</strong></td>
<td><strong>20.82%</strong></td>
</tr>
<tr>
<td>Blog 07</td>
<td>Most Improvement at Blog07 [13]</td>
<td>15.90%</td>
<td>21.60%</td>
<td>8.60%</td>
</tr>
<tr>
<td></td>
<td>M. Zhang et al. improvement [3]</td>
<td>28.10%</td>
<td>40.30%</td>
<td>19.90%</td>
</tr>
<tr>
<td></td>
<td>Our Relevant Baseline</td>
<td>0.2603</td>
<td>0.438</td>
<td>0.3131</td>
</tr>
<tr>
<td></td>
<td>Our Framework</td>
<td>0.3484</td>
<td>0.628</td>
<td>0.3869</td>
</tr>
<tr>
<td></td>
<td>Our framework’s improvement</td>
<td><strong>33.84%</strong></td>
<td><strong>43.38%</strong></td>
<td><strong>23.57%</strong></td>
</tr>
<tr>
<td>Blog 08 new topics</td>
<td>Most Improvement at Blog08 [14]</td>
<td><strong>31.60%</strong></td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Secondary best Improvement at Blog08</td>
<td>14.75%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Our Relevant Baseline</td>
<td>0.2818</td>
<td>0.498</td>
<td>0.3451</td>
</tr>
<tr>
<td></td>
<td>Our Framework</td>
<td>0.3296</td>
<td>0.6196</td>
<td>0.3789</td>
</tr>
<tr>
<td></td>
<td>Our framework's improvement</td>
<td>16.96%</td>
<td>24.42%</td>
<td>9.79%</td>
</tr>
</tbody>
</table>
In this paper...

1. We propose a novel framework to measure opinion relevance in the latent topic space.

2. It is essentially a general framework for opinion retrieval and can be used by most current solutions to further enhance their performance.

3. Our framework is domain-independent, thus fits for different opinion retrieval environments.

4. The effectiveness and advantage of our framework were justified by experimental results on TREC blog datasets.
Q&A

Many thanks!


