



Query-driven Opinion Summarization

Kam-Fai Wong

Department of Systems Engineering & Engineering
Management

The Chinese University of Hong Kong

Trends in WWW



香港中文大學工程學院
Faculty of Engineering
The Chinese University of Hong Kong

Content		
Read Only	Creation	Management
Web 1.0		
Web 2.0		
Web 3.0		

Web 3.0 Core Tech



香港中文大學工程學院
Faculty of Engineering
The Chinese University of Hong Kong

Social Networks

Integrated network (3-in-1)

Semantic Net

Ubiquitous and Mobile Computing

Internet Of Things (IOT)

Cloud Computing (IaaS, PaaS and SaaS)

Next Generation Internet (NGI, IPv6)

社會3.0現象



香港中文大學工程學院
Faculty of Engineering
The Chinese University of Hong Kong

- e-Engagement (people-centric)
- N-Generation
- Crowdsourcing
- Data is King -> People (relationship) is King
- Multi-modal
- Monetization

Outline


-  Introduction
-  Weighting Scheme
-  Query-driven Opinion Summarization
-  Evaluation
-  Conclusion & Future Works

Introduction

- With the explosion in the amount of commentaries on current issues and personal views expressed in weblogs, microblog on the Internet, there is a need to provide users a summary of opinions. [Kim et al. 2010]

Dell Inspiron 1545 - Pentium 2 GHz - 15.6 " - 3 GB Ram - 250 GB HD

[Overview](#) - [Compare prices](#) - [Reviews](#) - [Technical specifications](#) - [Similar items](#)

 **\$494 new, \$331 used** from [15 sellers](#)

★★★★★ 491 reviews

reviews

Summary - Based on 491 reviews

1 2 3 4 stars 5 stars

"Amazing laptop at an amazing price!"
"Great value, lightweight, and decent battery time."
"I love my laptop and it is easy to use."

"Battery is awful lasts for about 1 hour."
"Competitive pricing and a good value."
"Very durable laptop, great features for the price."

Dell Inspiron 1545 review
★★★★☆ By Chris Jager - Sep 22, 2009 - Editorial review - PC Advisor
The Dell Inspiron 1545 notebook is an affordable all-purpose notebook with a 15.6in widescreen LCD. While it's unlikely to turn heads with its pedestrian style, this Dell Inspiron 1545 laptop remains a perfectly serviceable notebook that punches well above its weight. (We use the term figuratively, as it is far from a petite notebook.) Weighing in at around 3kg and measuring 374x25.938mm, the Dell Inspiron 1545 is one of the bigger entry level notebooks on the market. This makes it a bit of a pain to lug around, but it will make an adequate desktop replacement - provided you're not into gaming. The main benefit of this added real-estate is a 15.6in screen with a native resolution of 1366x768. The display did a good job during movie playback, with excellent viewing angles and minimal reflective glare. While the inbuilt speakers are a little on the weak side, they're more than adequate for a notebook in this price range.

Business IT reviews and advice
Laptops reviews and buying advice
... [Read full review](#)

Introduction

- Aspect-based opinion summarization (AOS)
 - to divide input texts (mostly are review data) into aspects (features) and generate summaries of each aspect. [Liu et al. 2005]

e.g. Toshiba Satellite L655-S5158



- Query-driven opinion summarization (QOS)
 - to extract an informative summary of opinion expressions about a given query (also referred as topical opinion), as found in a document collection. [Dang 2008]
- e.g. “What complaints on YouTube do users have?”

Introduction

- Differences:
 - AOS is almost about review-type data, the aspects are limited to a list of predefined or labeled aspects for a same topic (product); while QOS concerns more on user's preference, and the query might only focus on one of the multiple topics presented in the related documents.

e.g. to summarize the negative opinions on YouTube from a number of articles on Internet Service.
 - In AOS, sentiment words are mostly domain-specific and the amount is fixed; while in QOS, general domain sentiment words will occur frequently across multiple topics.

e.g. Good vs. Delicious

Introduction

- One of the fundamental problems in QOS is how to effectively represent and measure topical opinion so as to precisely select the sentences with salient opinion expression.
- Existing methods:
 - Three-step approach :
 1. identify relevant text segments (e.g. sentences or passages) to the query from the blogs;
 2. re-rank the set of relevant segments by taking sentiment classification into consideration;
 3. select segments with high ranking and remove redundant text segments.

Most participants in TAC2008 adopted three-step approach. [V. Varma et al. 2008], [Razmara and Kosseim 2008], [Li et al. 2008],[Seki 2008], [Balahur et al. 2008]

Introduction

- Non-Three-step approach:
 - Stoyanov and Cardie [2008] proposed coreference resolution techniques to investigate the linkage between holders and the specific topic.

Limitations: opinion identification, opinion weighting and ranking opinions were not taken into consideration.

- Li et al. [2010] introduced word pair to express topical opinion and utilized a combination between the *tf-idf* values of topic and sentiment word to denote the weight of topical opinion.

$$w_{ij}^k = \frac{1}{|d_k|} \sum_{p_{ij} \in s_l \in d_k} [\lambda \cdot rel(t_i, s_l) + (1 - \lambda) opn(o_j, s_l)]$$

Limitations:

Domain-specific sentiment words will suffer from the weighting scheme of *tf*.

A unified parameter λ is inadequate to denote the variant associations between

Introduction

- Our Method:
 - Utilize word pair to represent topical opinion.
 - Measure the topical opinion by simultaneously considering the subjectivity of the topic word and the *local* relevance of the sentiment word.
 - Compute Pointwise Mutual Information(PMI) between sentiment word and its associated topic within a pair to measure the topical opinion in each individual word pair.
 - Implement the weighted topical opinions into a graph model for sentence ranking and MMR method to generate query-driven summary.

Outline



Introduction



Weighting Scheme



Query-driven Opinion Summarization



Evaluation



Conclusion & Future Works

Formal Definition

- Given a document set $D=\{d_1, d_2, d_3, \dots, d_n\}$, that includes a set of sentences $S=\{s_1, s_2, s_3, \dots, s_N\}$, and a specific query $Q=\{q_1, q_2, q_3, \dots, q_z\}$, where $q_1, q_2, q_3, \dots, q_z$ are query keywords.
- In addition, we construct a sentiment word lexicon V_o and a topic word lexicon V_t .
- We utilize the structure of *Query-sentiment word pair* p_{ij} to denote the topical opinion, which consists of two elements, one is from V_t , and the other one is from V_o . [Li et al, 2010]

$$p_{ij} = \{ \langle t_i, o_j \rangle \mid t_i \in V_t, o_j \in V_o \}$$

e.g. we can extract the word pair $\langle \text{Battery}, \text{awful} \rangle$ from the sentence “Battery is awful lasts for about 1 hour”.

Topical Opinion Weighting

- We measure topical opinion based on the following assumptions:
 - topic word t_1 is more important than topic word t_2 when there are more comments or opinions on t_1 than t_2 .
 - sentiment word o_1 can be regarded as domain-specific sentiment word due to different associated targets.
 - the associations between topic and sentiment words in different word pairs vary a lot. [Kim et al. 2009]

Topical Opinion Weighting

- We measure topical opinion in 2 stages:
 - we first measure both topic word and sentiment word by computing the gain in selecting a sentence containing the word.
 - Based on pairwise representation, we weigh topical opinion by computing the PMI between the target and sentiment words within a pair.
- Assume that a term t follows the distribution $P_D(t)$ on the whole set of words, and it also follows another distribution $P_{S(t)}(t)$ on the set of sentences including t . The higher deviation of $P_{S(t)}(t)$ from $P_D(t)$, the higher the information content of t is.

$$Inf = -\log P_{S(t)}(t)/P_D(t)$$

Topical Opinion Weighting

- From an information theoretic point of view, not all the sentiment words appear randomly, e.g. “good, bad”.
- *Inf* considers only the divergence between two distributions, which will generate a bias on domain-specific sentiment words.

Topical Opinion Weighting

- We take the advantage of pairwise representation to convert all sentiment words as domain-specific sentiment, and apply PMI to assess the subjective of a target and the local relevance of a sentiment word. [Turney 2002]
- The PMI between t_i and o_j within p_{ij} can be computed by:

$$I(X_{t_i}; X_{o_j}) = \sum_{X_{t_i}=0,1} \sum_{X_{o_j}=0,1} \log \frac{p(X_{t_i}, X_{o_j})}{p(X_{t_i})p(X_{o_j})}$$

where X_{t_i} and X_{o_j} denote whether t_i , o_j appear in the sentence

$$\begin{aligned} p(X_{t_i} = 1) &= c(X_{t_i} = 1) / C(N) & p(X_{t_i} = 1, X_{o_j} = 0) &= \frac{c(X_{t_i} = 1) - c(X_{t_i} = 1, X_{o_j} = 1)}{C(N)} \\ p(X_{t_i} = 0) &= 1 - p(X_{t_i} = 1) & p(X_{t_i} = 0, X_{o_j} = 1) &= \frac{c(X_{o_j} = 1) - c(X_{t_i} = 1, X_{o_j} = 1)}{C(N)} \\ p(X_{o_j} = 1) &= c(X_{o_j} = 1) / C(N) & p(X_{t_i} = 0, X_{o_j} = 0) &= 1 - p(X_{t_i} = 1, X_{o_j} = 1) \\ p(X_{o_j} = 0) &= 1 - p(X_{o_j} = 1) & & - p(X_{t_i} = 0, X_{o_j} = 1) \\ p(X_{t_i} = 1, X_{o_j} = 1) &= \frac{c(X_{t_i} = 1, X_{o_j} = 1)}{C(N)} & & - p(X_{t_i} = 1, X_{o_j} = 0) \end{aligned}$$

Topical Opinion Weighting

- According to the assumptions, we can assign the topic word and sentiment word as:

$$Weight(t_i) = \mu_{t_i} \cdot Inf(t_i) \quad \text{and} \quad Weight(o_j) = \mu_{o_j} \cdot Inf(o_j)$$

$$\text{where, } \mu_{t_i}(o|t_i) = \sum_{o \in V_o} p(X_{t_i}, X_o) \cdot I(X_{t_i}; X_o)$$

$$\mu_{o_j}(t|o_j) = \sum_{t \in V_t} p(X_t, X_{o_j}) \cdot I(X_{o_j}; X_t)$$

- We finally add the associative score to each word pair and compute the weight of a topical opinion as:

$$w_{p_{ij}} = \lambda_{ij}(t_i, o_j) \cdot [Weight(t_i) + Weight(o_j)]$$

$$\text{where } \lambda_{ij}(t_i, o_j) = I(X_{t_i}; X_{o_j})$$

Outline



Introduction



Weighting Scheme



Query-driven Opinion Summarization



Evaluation



Conclusion & Future Works

Sentence Ranking

- To generate a summary for a specific query, we first select a set of sentences with the topical opinions.
- Graph-based ranking algorithms, such as HITS or PageRank, have been traditionally and successfully used in citation analysis, retrieval, summarization. [Erkan et al. 2004]
- Intuitively, sentences containing more word pairs with the topical opinions should achieve a relatively higher ranking.

Sentence Ranking

- Based on the PageRank model, we define a graph with nodes representing relevant sentences and edges connecting 2 sentences sharing a common word pair.
- We then score all the sentences based on the expected probability of a random walker visiting each sentence.
- The jumping probability $P(s_u|s_v)$ from node s_v to node s_u is given by:

$$P(s_u|s_v) = \frac{\text{sim}(s_v, s_u)}{\sum_{u \in S \setminus v} \text{sim}(s_v, s_u)}$$

where

$$\text{sim}(s_u, s_v) = \frac{\sum_{p_{ij} \in s_u, s_v} w_{p_{ij}}^{s_u} \cdot w_{p_{ij}}^{s_v}}{\sqrt{\sum_{p_{ij} \in s_u} (w_{p_{ij}}^{s_u})^2} \times \sqrt{\sum_{p_{ij} \in s_v} (w_{p_{ij}}^{s_v})^2}}$$

Sentence Ranking

- All sentences are initialized equally. In each iteration $T+1$, the scores are updated according to the scores in iteration T .

$$Score(s_u)^{T+1} = \gamma \sum_{v \neq u} Score(s_v)^T \cdot P(s_u|s_v) + (1 - \gamma) \cdot sim(s_u|Q)$$

- Iteration is terminated when the maximum difference between the scores computed for two successive iterations is lower than a given threshold (empirically setting as 0.00001). [Li et al. 2009]
- Finally, the sentences are ranked by the scores.

QOS

- We adopt maximal marginal relevance (MMR) method to generate the summary by incrementally adding the top ranked sentences into the answer set. [Carbonell and Goldstein 1998]

$$MMR = Arg \max_{s_u \in R \setminus S'} \left[\theta (sim(s_u|Q)) - (1 - \theta) \max_{s_v \in S'} sim(s_u, s_v) \right]$$

- R is the ranked list of sentences retrieved in the previous step. We set a relevant threshold, below which it will not be regarded as candidate sentences.
- The parameter θ lying between $[0,1]$ controls the relative importance given to relevance versus redundancy. In our experiments we set $\theta=0.5$.

Outline



Introduction



Weighting Scheme



Query-driven Opinion Summarization



Evaluation



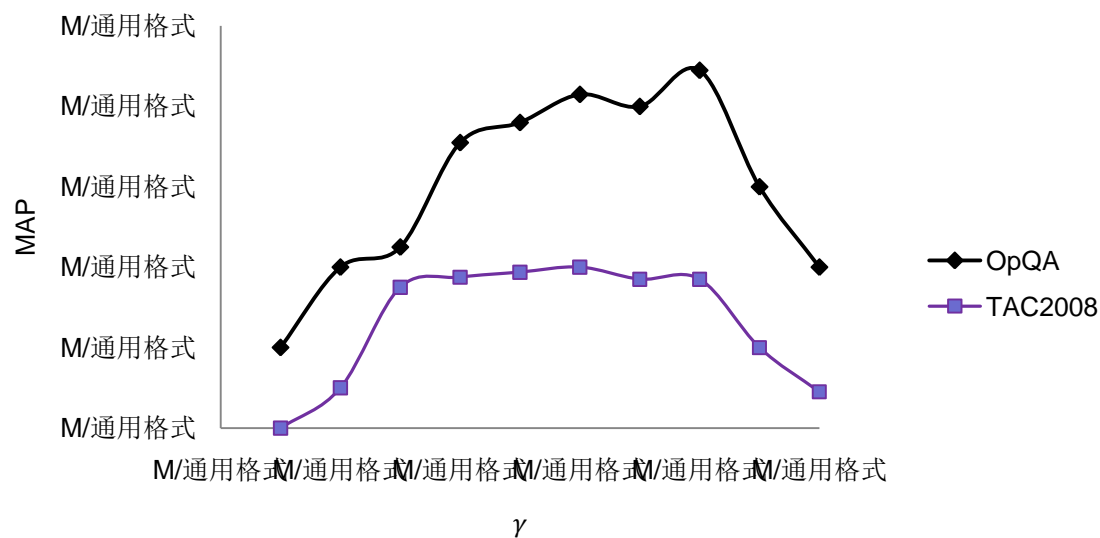
Conclusion & Future Works

Experiment Setting

- Datasets:
 - TAC08 dataset is the benchmark data set for the query-driven opinion summarization track in the Text Analysis Conference 2008 (TAC2008), which contains a total number of 2500 documents and 87 opinion queries. [Dang, 2008]
 - The Opinion Question Answering (OpQA) corpus consists of 98 documents appeared in the world press and 30 queries. [Wilson, et al 2005]
- Sentiment Lexicon:
 - We use SentiWordNet as the sentiment lexicon, which consists of 4800 negative sentiment words and 2290 positive sentiment words.
- Topic Word Collection:
 - The dictionary-based method
 - The web-based method

Parameter Tuning

- Experimental parameter tuning (γ)



— $\gamma=0.8$

Experiment 1

- In our evaluation, we first test the performance of our proposed weighting scheme for measuring topical opinion.
- Methods for comparison:
 - *tf-idf*
 - WordNet: applied the maximum value of a sentiment word in SentiWordNet Lexicon as the weight of sentiment word.
 - GOSM: proposed to represent topical opinion by word pair, and utilized *tf-idf* to weigh topical opinion.
 - PPM: our proposed method.
- Experimental Metrics:
 - MAP: Mean Average Precision
 - Rpre: R-precision
 - P@10

Experimental Result 1

- Comparison of different weighting schema on TAC08 and OpQA, and the best result in each column is highlighted.

Table 1: Comparison of sentence ranking on OpQA and TAC2008 datasets

Dataset	Probability	Metrics		
		MAP	R-Prec	P@10
OpQA	GOSM	0.212	0.233	0.408
	<i>tf-idf</i>	0.208	0.230	0.397
	WordNet	0.195	0.214	0.366
	PPM	0.229	0.245	0.421
TAC 2008	GOSM	0.177	0.206	0.361
	<i>tf-idf</i>	0.175	0.198	0.354
	WordNet	0.166	0.196	0.322
	PPM	0.180	0.212	0.369

Experiment 2

- Different approaches for QOS for comparison:
 - Baseline 1: This model was achieved the best run in TAC2008 opinion summarization task. [Varma, et al., 2008]
 - Baseline 2: This model was achieved 10% improvement over the best run in TAC2008 Opinion QA track. We modified this model to deal with QOS. [Li, et al., 2009]
 - OPM: similar with Baseline 2, but use PageRank model for sentence ranking instead.
 - GOSM: This model adopted pairwise representation of topical opinion. We re-designed GOSM to deal with QOS. [Li, et al., 2010]
 - PPM: our proposed approaches.
- Experimental Metrics:
 - Precision
 - Recall
 - F-value:

Experimental Result 2

- Comparison of different approaches for opinion summarization on TAC08 and OpQA datasets, and the best $F(3)$ is highlighted.

Data set	Approaches	Measurements		
		Precision	Recall	F(3)
OpQA	Baseline 1	0.280	0.356	0.325
	Baseline 2	0.274	0.368	0.336
	OPM	0.281	0.354	0.325
	GOSM	0.286	0.360	0.300
	PPM	0.276	0.375	0.355
TAC 2008	Baseline 1	0.101	0.217	0.186
	Baseline 2	0.102	0.256	0.205
	OPM	0.113	0.245	0.198
	GOSM	0.102	0.242	0.196
	PPM	0.103	0.268	0.213

Table 2. Comparison of TOS on OpQA and TAC2008 datasets.

Outline



Introduction



Weighting Scheme



Query-driven Opinion Summarization



Evaluation



Conclusion & Future Works

Conclusion

- We utilize pairwise representation to denote topical opinion.
- A weighting scheme has been proposed to measure the topical opinion by simultaneously considering the subjectivity of the topic word and the *local* relevance of the sentiment word.
- Weighted topical opinions were implemented into a graph model for sentence ranking and MMR method to generate query-driven summary which performs well on TAC2008 and OpQA datasets.

Future works

- Need of techniques of other research areas
 - Deeper NLP e.g., discourse analysis, dependency parser, may help to understand the meaning of opinion so as to improve the accuracy.
- Need of Standardized Data Set and Evaluation
 - Current published data set are depending on their own purpose and lack of widely used dataset. [Hu and Liu 2004], [Kim and Zhai 2009], [Ganesan et al. 2010]
 - Lack of evaluation measures which cover entire opinion summarization steps is another issue.

References

- H. Dang, 2008. Overview of the TAC 2008 opinion question answering and summarization tasks. In *TAC 2008*.
- G. Erkan and D. R. Radev. 2004. Lexpagerank: Prestige in multi-document text summarization. In *EMNLP 2004*.
- B. Ernsting, W. Weerkamp, and M. de Rijke. 2007. Language modeling approaches to blog post and feed finding. In *TREC 2007*.
- A. Esuli and F. Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*.
- B. He, C. Macdonald, J. He, I.h Ounis. 2008. An effective statistical approach to blog post opinion retrieval, In *CIKM 2008*.
- M. Hu, B. Liu, 2004. Mining and summarizing customer reviews, Proceedings of the tenth ACM SIGKDD 2004.
- J. Kim, J. Li, and J. Lee. 2009. Discovering the discriminative views: Measuring term weights for sentiment analysis. *ACL-IJCNLP 2009*.
- B. Li, L. Zhou, S. Feng and KF Wong. 2010. A unified graph model for sentence-based opinion retrieval, In *ACL 2010*.

References

- F. Li, Y. Tang, M. Huang, and X. Zhu. 2009. Answering opinion questions with random walks on graphs. In *ACL 2009*.
- B. Liu, M. Hu, and J. Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *WWW 2005*.
- B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*
- V. Stoyanov, C. Cardie, 2006. Toward opinion summarization: linking the sources, *Proceedings of the Workshop on Sentiment and Subjectivity in Text*.
- P. Turney, 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *ACL 2002*.
- V. Varma, P. Pingali, R. Katragadda, S. Krishna, S. Ganesh, K. Sarvabhotla, H. Garapati, H. Gopisetty, V.B. Reddy, K. Reddy, P. Bysani, R. Bharadwaj, 2008. IIIT Hyderabad at TAC 2008, In *TAC2008*.
- T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *EMNLP 2005*.



Q&A