Mining Opinions from the Web: Beyond Relevance Retrieval

Lun-Wei Ku and Hsin-Hsi Chen*
Department of Computer Science and Information Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, Taiwan. E-mail: lwku@nlg.csie.ntu.edu.tw and hhchen@csie.ntu.edu.tw

Documents discussing public affairs, common themes, interesting products, and so on, are reported and distributed on the Web. Positive and negative opinions embedded in documents are useful references and feedbacks for governments to improve their services, for companies to market their products, and for customers to purchase their objects. Web opinion mining aims to extract, summarize, and track various aspects of subjective information on the Web. Mining subjective information enables traditional information retrieval (IR) systems to retrieve more data from human viewpoints and provide information with finer granularity. Opinion extraction identifies opinion holders, extracts the relevant opinion sentences, and decides their polarities. Opinion summarization recognizes the major events embedded in documents and summarizes the supportive and the nonsupportive evidence. Opinion tracking captures subjective information from various genres and monitors the developments of opinions from spatial and temporal dimensions. To demonstrate and evaluate the proposed opinion mining algorithms, news and bloggers' articles are adopted. Documents in the evaluation corpora are tagged in different granularities from words, sentences to documents. In the experiments, positive and negative sentiment words and their weights are mined on the basis of Chinese word structures. The f-measure is 73.18% and 63.75% for verbs and nouns, respectively. Utilizing the sentiment words mined together with topical words, we achieve f-measure 62.16% at the sentence level and 74.37% at the document level.

Introduction

Documents discussing public affairs, common themes, interesting products, and other topics are reported and distributed on the Web. Watching specific information sources and summarizing the newly discovered opinions are important for governments to improve their services, for companies to market their products, and for customers to purchase their objects. Web opinion mining aims to extract, summarize, and track various aspects of subjective information on the Web. These tasks are different from traditional information retrieval and extraction. Conventional information retrieval only identifies which document is relevant to a given topic; it does not identify whether it is positive, negative, or neutral about the topic. In addition, conventional information extraction only recognizes the major components such as named entities and their relationships; it does not report the supportive and the nonsupportive evidence.

On the other hand, Web opinion mining can enhance conventional information retrieval in several ways. Although queries are usually objective, an information retrieval system can still be active to provide subjective information on the Web. For example, when query terms are named entities, Web users may want to know the opinions toward these entities. From the information visualization, facts, opinions, or both can be presented. Separating the retrieval results into objective and subjective types gives Web users an alternative way to find their interesting information. Opinion information processing is the first step to achieve the goals stated.

Opinion information processing can be discussed from different viewpoints, such as granularities of information, information sources, and methodologies. Documents, sentences, and words express three different levels of subjective information for opinion mining. Wiebe, Wilson, and Bell (2001) recognize opinion documents. Pang, Lee, and Vaithyanathan (2002) classify documents by overall sentiments instead of topics. Riloff and Wiebe (2003) distinguish subjective sentences from objective ones. Kim and Hovy (2004) present a sentiment classifier for English words and sentences by utilizing thesauri. Riloff, Wiebe, and Wilson (2003) propose a method to extract subjective nouns by using patterns. However, a template-based approach requires a professionally annotated corpus for learning, and words in thesauri are not always consistent in sentiment.

Reviews are often adopted for opinionated information processing because of their practical applications. Dave, Lawrence, and Pennock (2003) extract opinions from product reviews. Hu and Liu (2004) propose opinion summarization of products; Liu, Hu and Cheng (2005) illustrate

*Corresponding author

Accepted January 4, 2007

• Published online 3 August 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20630

opinion summarization in bar graph style. Bai, Padman, and Airoldi (2005) categorize movie reviews by opinion polarities. However, reviews express only one genre of opinions; e.g., they are usually toward Computer, Communication and Control (3C) products, books or movies. There are various sources of subjective information on the Web. News and blogs are two of them. Extracting opinions in news and blogs is different from that in domain-dependent reviews. Rich topics are embedded in news and blogs, so that topic detection is critical to expel nonrelevant sentences (Ku, Lee, Wu, & Chen, 2005a). In addition, single document summarization is not enough. Knowing how to summarize points of view from various communities and individuals is indispensable.

Several approaches have been proposed for opinion extraction. Pang, Lee, and Vaithyanathan (2002) show that machine learning approaches on sentiment classification do not perform as well as those on traditional topic-based categorization at the document level. Information extraction technologies (Cardie, Wiebe, Wilson, & Litman, 2003) have also been explored. Hiroshi, Tetsuya, and Hideo (2004) regard the extraction of sentiment units as a kind of translation and adopt machine translation technology. A statistical model is used for the sentiment word analysis (Takamura, Inui, & Okumura, 2005). The results for various metrics and heuristics depend on the testing situations.

In this paper, an opinion mining system that provides relevancy in finer granularity is proposed. We deal with opinion mining of different information sources (news and blogs) in different languages (Chinese and English) at different levels (words, sentences, and documents) and in different applications (extraction, summarization, and tracking). Opinion Extraction Algorithms first discusses opinion extraction algorithms. Evaluation Design then introduces the evaluation design, and Evaluations of Opinion Extraction shows the performance of opinion extraction algorithms on the experimental corpora. Opinion Summarization and Opinion Tracking System present the algorithms for summarization and tracking. In practical applications, information needs are not always posed clearly beforehand. In addition to distinguishing between positive and negative polarity, identifying which events correlated with which opinions is important. A topic detection algorithm is introduced to capture the main concepts embedded implicitly in a document set and is applied to retrieve relevant sentences for opinion summarization. In opinion summarization, we retrieve all the relevant sentences related to the major topic from the document set, determine the opinion polarity of each relevant sentence, and finally summarize positive sentences and negative sentences, respectively. Opinion tracking monitors the developments of opinions from spatial and temporal dimensions.

Opinion Extraction Algorithms

Opinion extraction is the basis of mining opinions. Its aim is to extract opinion evidence from words, sentences, and documents, and then to identify their polarities. Documents collected are clues to decide the major opinions of the public on a specific topic. Likewise, sentences serve as evidence when identifying the major opinions of a document and words are considered when judging the qualification of an opinion sentence. In the following sections, algorithms are explained in detail from the bottom level up to the top level.

TABLE 1. Qualified seeds.

Dictionary	Positive	Negative
GI	2,333	5,830
CNSD	431	1,948
Total	2,764	7,778

Word Level

Sentiment words are employed to compute the tendency of a sentence, and then a document. To detect sentiment words in Chinese documents, a Chinese sentiment dictionary is indispensable. However, a small dictionary may suffer from the problem of coverage. A method to learn sentiment words and their strengths from multiple resources is developed in this section.

First, two sets of sentiment words, General Inquirer[1] (GI) and Chinese Network Sentiment Dictionary[2] (CNSD), are col-lected. The former is in English and we translate those words into Chinese. The latter, whose sentiment words are collected from the Internet, is in Chinese. Table 1 shows the statistics of the revised dictionaries. Words from these two resources form the "seed vocabulary" in the opinion dictionary. These qualified seeds are collected in National Taiwan University Sentiment Dictionary (NTUSD) and available from uniform resource locator (URL) http://nlg18.csie.ntu.edu.tw:8080/opinion/index.html.

Then, the seed vocabulary is enlarged by consulting two thesauri: tong2yi4ci2ci2lin2 (abbreviated Cilin) (Mei, Zhu, Gao, & Yin, 1982) and the Academia Sinica Bilingual Ontological Wordnet[3] (abbreviated BOW). Cilin is composed of 12 large categories, 94 middle categories, 1,428 small categories, and 3,925 word clusters. BOW is a Chinese thesaurus with a similar structure to WordNet's.[4] However, words in the same clusters may not always have the same opinion tendency. For example, 「寬恕」(forgive: positive) and 「姑息」(appease: negative) are in the same synonym set (synset). Nevertheless, they have opposite opinion tendencies. How to distinguish words with different polarities within the same cluster/synset is the major issue of using thesauri to expand the seed vocabulary and is addressed later.

It is postulated that the meaning of a Chinese sentiment word is a function of the composite Chinese characters. This is exactly how people read an ideogram when they encounter a new word. A sentiment score is then defined for a Chinese

[1] http://www.wjh.harvard.edu/~inquirer/

[2] http://134.208.10.186/WBB/EMOTION_KEYWORD/Atx_emtwordP.htm

[3] http://bow.sinica.edu.tw/

[4] http://wordnet.princeton.edu/

DOI: 10.1002/asi

word by the following formulas. The equations not only tell us the opinion tendency of an unknown word, but also suggest its strength. Moreover, using these equations, synonyms of different polarities are distinguishable by their scores while applying thesaurus expansion. The discussion begins with the definition of the formula of Chinese characters.

$$P_{c_i} = \frac{fp_{c_i}}{fp_{c_i} + fn_{c_i}} \tag{1}$$

$$N_{c_i} = \frac{fn_{c_i}}{fp_{c_i} + fn_{c_i}} \tag{2}$$

where fp_{c_i} and fn_{c_i} denote the frequencies of a character c_i in the positive and negative words, respectively.

Formulas (1) and (2) utilize the percentage of a character in positive/negative words to show its sentiment tendency. However, there are more negative words than positive ones in the "seed vocabulary." Hence, the frequency of a character in a positive word may tend to be smaller than that in a negative word. That is unfair for learning, so that formulas (1) and (2) are normalized into formulas (3) and (4), respectively.

$$P_{c_i} = \frac{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j}}{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j} + fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}} \tag{3}$$

$$N_{c_i} = \frac{fn_{c_i} / \sum_{j=1}^{n} fn_{c_j}}{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j} + fn_{c_j} / \sum_{j=1}^{m} fn_{c_j}} \tag{4}$$

where P_{ci} and N_{ci} denote the weights of c_i as positive and negative characters, respectively; n and m denote total number of unique characters in positive and negative words, respectively. The difference of P_{ci} and N_{ci}, i.e., $P_{ci} - N_{ci}$ in formula (5), determines the sentiment tendency of character c_i. If it is a positive value, this character appears in positive Chinese words more often than in negative ones, and vice versa. A value close to 0 means that it is not a sentiment character or it is a neutral sentiment character.

$$S_{c_i} = (P_{c_i} - N_{c_i}) \tag{5}$$

Formula (6) defines a sentiment degree of a Chinese word w as the average of the sentiment scores of the composing characters $c_1, c_2, \ldots, c_p$.

$$S_w = \frac{1}{p} \times \sum_{j=1}^{p} S_{c_j} \tag{6}$$

If the sentiment score of a word is positive, it is likely to be a positive sentiment word, and vice versa. A word with a sentiment score close to 0 is possibly neutral or nonsentiment. By considering sentiment scores of words, sentiment words can be detected. With the sentiment words extracted, opinion tendencies of sentences and documents can be decided.

Sentence Level

When judging the opinion polarity of a sentence, three factors are considered: sentiment words, negation operators, and opinion holders. Every sentiment word has its own sentiment score. If a sentence consists of more positive sentiments than negative sentiments, it must reveal something good, and vice versa. However, a negation operator, such as *not* or *never*, may totally change the sentiment tendency of a sentiment word. Therefore, when a negation operator appears together with a sentiment word, the opinion score of the sentiment word $\boldsymbol{S}$ is changed to $-\boldsymbol{S}$ to keep the strength but reverse the tendency. A total of 41 negation words are employed in this paper. In addition to the overt negation operators such as *not* and *never*, other terms, such as *unable*, *impossible*, *without*, and *unlikely*, are collected in the list. Opinion holders are also considered at sentence level, but the way they influence opinions has not yet been investigated. As a result, they are weighted equally at first. The algorithm for calculating the opinion score of one sentence is shown in the following.

Algorithm: *Opinion Sentence Extraction*

```
1. For every sentence $p$
2.   For every sentiment word in $p$
3.     If a negation operator appears before, then reverse
         the sentiment tendency.
4. Decide the opinionated tendency of $p$ by the function of
     sentiment words and the opinion holder as follows.
```

$$S_p = S_{opinion\text{-}holder} \times \sum_{j=1}^{n} S_{w_j} \tag{7}$$

Where S_p, $S_{opinion\text{-}holder}$, and S_{wj} are the sentiment score of sentence p, the weight of *opinion holder*, and the sentiment score of word w_j, respectively, and n is the total number of sentiment words in p.

Document Level

Whether a document contains opinions depends on the number of opinion sentences it has and how strong they are. Therefore, the sentiment scores of sentences are summed up to be the final sentiment score of one document. The following is an algorithm for calculating the opinion score of one document.

Algorithm: *Opinion Document Extraction*

```
1. For every document $d$
2. Decide the opinion tendency of $d$ by the function of the
     opinion tendencies of sentences insides $d$ as follows.
```

$$S_d = \sum_{p=1}^{m} S_p \tag{8}$$

Where S_d and S_p are sentiment scores of document d and sentence p, and m is the amount of evidence. If the topic is *anti* type, reverse the sentiment type.

```
<topic>

<number> CIRB010TopicZH027 </number>

<title>反美濃水庫興建(Anti-Meinung Dam Construction) </title>

<concepts>

美濃(Meinung),水庫(Dam),美濃水庫(Meinung Dam),反水庫(Anti-

Dam),抗爭(resist),興建(build),斷層(fault),☒染(pollution), 地質 (geology), 工業

(industry), 環境(enviroment),安全(safety),水資源(water resources), 生態(ecology),

水質(water quality), 替代方案 (displace program)

</concepts>

</topic>
```

FIG. 1. A NTCIR topic description.

Evaluation Design

Three sources of information are collected for the experiments: Test Retrieval Conference (TREC)[5] corpus, NACSIS Test Collection for IR (NTCIR)[6] corpus for Web news, and BLOG corpus from weblogs (BLOGs) corpus from the Web. TREC and NTCIR are two of three major information retrieval evaluation forums in the world. TREC corpus is in English, while the other two are in Chinese. NTCIR corpus is employed to study the fundamental opinion extraction algorithms. To discuss the characteristics of different information sources on opinion extraction, experiments using NTCIR corpus and BLOG corpus are compared. These three corpora are used for opinion summarization to deal with the multilingual issue emerging from the combination of information sources in different languages.

Corpus Preparation

For opinion extraction, NTCIR corpus is used. Chen and Chen (2001) developed a test collection CIRB010 for Chinese information retrieval in NTCIR 2. The test collection consists of 50 topics and 132,173 Chinese documents. Each topic in CIRB010 test collection is in TREC style.

Of the 50 topics in CIRB010, six opinion topics are chosen for experiments on opinion extraction. The topics and the corresponding number of relevant documents are shown in Table 2.

TABLE 2. Opinion topics in NTCIR corpus.

Topic ID	Topic title	Total documents
ZH021	Civil ID Card	37
ZH024	The Abolishment of Joint College Entrance Examination	55
ZH026	The Chinese-English Phonetic Transcription System	30
ZH027	Anti-Meinung Dam Construction	14
ZH028	Hewing Down of Chinese Junipers in Chilan	23
ZH036	Surrogate Mother	33

As an example, topic ZH027 related to environmental protection is presented. For simplicity, only <title> denoting a request subject and <concepts> consisting of relevant keywords are shown in Figure 1.

A total of 192 documents relevant to these six topics are chosen as training data in this paper. Annotators assign *positive, neutral,* and *negative* tags (<DOC_ATTITUDE>) to opinion documents. In addition to documents, the annotators assign polarities of sentences (<SEN_ATTITUDE>): *supportive*, *neutral*, and *nonsupportive*.

Furthermore, the annotators add the tags *positive, neutral, negative keyword* (<SENTIMENT_KW>), and *opinion operator* (<OPINION_OPR>) to the critical words in the passages. *Positive keyword*s like "成功" (succeed) and *negative keyword*s like "質疑" (suspect; challenge) are sentiment words that express positive and negative attitudes. In contrast, the *opinion operators* like "表示" (express) only signal opinions and do not indicate a clear sentiment tendency.

[5] http://trec.nist.gov/

[6] http://research.nii.ac.jp/ntcir/index-en.html

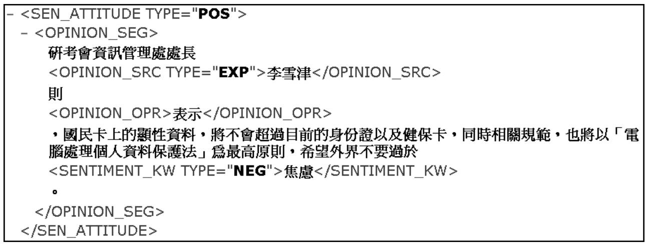
DOI: 10.1002/asi

TABLE 3. Tag descriptions.

Tag			
Level	Attribute	Value	Description
		<DOC_ATTITUDE></DOC_ATTITUDE>	
Document	TYPE	POS NEG NEU	Document attitude: Define the opinion polarity of the whole document
		<SEN_ATTITUDE></SEN_ATTITUDE>	
Sentence	TYPE	SUP NSP NEU	Sentence Attitude: Define the opinion polarity of one sentence
		<OPINION_SEG></OPINION_SEG>	
Sub-sentence	TYPE	PSV	Opinion segment: Define the scope of one opinion
		<OPINION_SRC></OPINION_SRC>	
Sub-sentence	TYPE	EXP IMP	Opinion source: Define the opinion holder of a specific opinion
		<SENTIMENT_KW></ SENTIMENT_KW >	
Word	TYPE	POS NEG NEU	Sentiment keyword: Define the opinion polarity of a single word.
		<OPINION_OPR></OPINION_OPR>	
Word	TYPE	PSV	Opinion operator: Define the keyword of expressing an opinion.

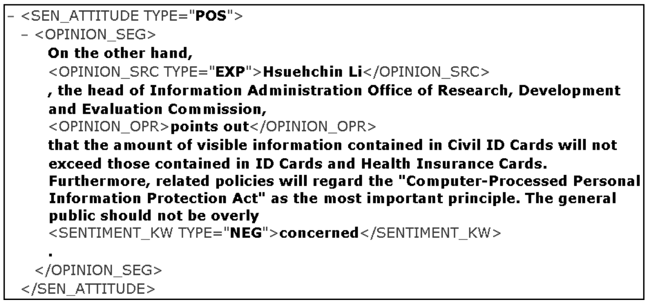
```
-<OPINION_SEG>
   <OPINION_SRC>A</OPINION_SRC>
   <OPINION_OPR>says</OPINION_OPR>
   that
   -<OPINION_SEG>
      <OPINION_SRC>B</OPINION_SRC>
      <OPINION_OPR>insists</OPINION_OPR>
      event C
   </OPINION_SEG>
   and
   -<OPINION_SEG>
      <OPINION_SRC>D</OPINION_SRC>
      <OPINION_OPR>disproves</OPINION_OPR>
      event C
   </OPINION_SEG>
</OPINION_SEG>
```

FIG. 2. A sample of nested tags.



```
- <SEN_ATTITUDE TYPE="POS">
  - <OPINION_SEG>
      研考會資訊管理處處長
      <OPINION_SRC TYPE="EXP">李雪津</OPINION_SRC>
      則
      <OPINION_OPR>表示</OPINION_OPR>
      ，國民卡上的顯性資料，將不會超過目前的身份證以及健保卡，同時相關規範，也將以「電
      腦處理個人資料保護法」爲最高原則，希望外界不要過於
      <SENTIMENT_KW TYPE="NEG">焦慮</SENTIMENT_KW>
      。
    </OPINION_SEG>
  </SEN_ATTITUDE>
```

FIG. 3. A Civil ID card example in Chinese.



```
- <SEN_ATTITUDE TYPE="POS">
  - <OPINION_SEG>
      On the other hand,
      <OPINION_SRC TYPE="EXP">Hsuehchin Li</OPINION_SRC>
      , the head of Information Administration Office of Research, Development
      and Evaluation Commission,
      <OPINION_OPR>points out</OPINION_OPR>
      that the amount of visible information contained in Civil ID Cards will not
      exceed those contained in ID Cards and Health Insurance Cards.
      Furthermore, related policies will regard the "Computer-Processed Personal
      Information Protection Act" as the most important principle. The general
      public should not be overly
      <SENTIMENT_KW TYPE="NEG">concerned</SENTIMENT_KW>
      .
    </OPINION_SEG>
  </SEN_ATTITUDE>
```

FIG. 4. A Civil ID card example in English.

Table 3 lists the annotation tags and their corresponding descriptions, with the following attribute values: POS: positive, NEG: negative, NEU: neutral, SUP: supportive, NSP: nonsupportive, PSV: preserved, EXP: explicit, and IMP: implicit.

Every element has an opening and a closing tag as the Extensible Markup Language (XML) language. For example, the pair <DOC_ATTITUDE> and </DOC_ATTITUDE> denotes document attitude. The tag <OPINION_SEG> is especially useful in dealing with multiperspective or opinion holder related issues. Consider the following example:

A says that B insists event C and D disproves event C. It is tagged as in Figure 2.

Nested relations of opinion holders are critical to identify the holding of opinions, that is, multiperspective issues. XML-like tags can easily represent nested relations. A Chinese and an English tagging example are illustrated in Figures 3 and 4, which show an opinion passage for topic

DOI: 10.1002/asi

TABLE 4. Corpora relevant to “animal cloning.”

Source	Quantity
TREC	25
NTCIR	17
BLOG	20

ZH021 in Chinese and in English, respectively. This topic concerns the personal privacy issue of a government policy. The opinion operator “表示” (point out) shows that this passage may be an opinion, and a negation “不要” (should not) that modifies a nonsupportive keyword “焦慮” (concerned) transforms a negative passage to a positive one. The opinion holder is “李雪津” (Hsuehchin Li). Documents of six topics are annotated using these tags for later experiments (Ku, Wu, Lee, & Chen, 2005b).

TREC, NTCIR, and BLOG corpora are used for opinion summarization. TREC corpus is from the TREC 2003 novelty track (Soboroff & Harman, 2003). There are 50 document sets in TREC novelty corpus, and each set contains 25 documents. All documents in the same set are relevant. To investigate opinion summarization from different information sources, a common topic among three corpora—say, animal cloning—is adopted. In TREC corpus, set 2 is “clone Dolly sheep.” It discusses the feasibility of the gene cloning and the perspectives of the authority. For NTCIR corpus, an additional topic, “animal cloning,” is selected from NTCIR 3 (Kishida et al., 2004). The blog is a new rising community for expressing opinions. To investigate the opinions in blogs, we retrieve documents from blog portals by the query “animal cloning.” Relevant documents from the two largest Chinese blog portals in Taiwan, YamBlog (http://blog.yam.com/) and WRETCH (http://www.wretch.cc/blog/), are collected. The numbers of documents relevant to “animal cloning” in the three different information sources are listed in Table 4.

Agreement Analysis

To evaluate the quality of the human-tagged corpora, the agreements of annotations are analyzed. Interannotator agreements are conducted at word, sentence, and document levels, then kappa values give a quantitative measure of tagging agreement.

TABLE 5. Agreement of annotators under strict metrics.

Annotators	A vs. B	B vs. C	C vs. A	Average
Percentages	78.64%	60.74%	66.47%	68.62%
All agree		54.06%		

TABLE 6. Agreement of annotators under lenient metrics.

Annotators	A vs. B	B vs. C	C vs. A	Average
Percentages	79.47%	62.05%	67.54%	69.69%
All agree		55.13%		

TABLE 7. Agreement of annotators at sentence level.

Annotators	A vs. B	B vs. C	C vs. A	Average
Percentages	73.06%	68.52%	59.67%	67.11%
All agree		52.19%		

TABLE 8. Agreement of annotators at document level.

Annotators	A vs. B	B vs. C	C vs. A	Average
Percentages	73.57%	68.86%	60.44%	67.62%
All agree		52.86%		

The tag of the smallest granularity, i.e., <SENTIMENT_KW>, is evaluated first. Because it is not cost effective to examine all the sentiments tagged by all annotators, only the words that are of parts of speech *noun, verb, adjective,* and *adverb,* and co-occur with one of the seeds (defined in Word Level) are sampled for agreement test. A total of 838 words were selected. The metrics of the interannotator agreement is shown in formula 9.

$$Agreement(A, B) = \frac{A \cap B}{samples} \tag{9}$$

Three annotators denoted A, B, and C examined the samples. Tables 5 and 6 show the agreement of the three annotators under strict and lenient metrics. Under lenient metrics, neutral sentiment words and positive sentiment words are considered to be in the same category. Strict metrics treats all three categories (positive, neutral, and negative) as distinct. The average agreements between two annotators under the two metrics are 68.62% and 69.69%, and the agreements among three annotators are 54.06% and 55.13%, respectively.

The largest category of samples is “nonsentiment” (400 out of 838 words, i.e., 47.73%). The agreements of two annotators (68.62% and 69.69%) are much larger than the baseline percentage (47.73%). In addition, the agreements of all annotators (54.60% and 55.13%) are still significantly higher. Here, we do not define a strength tag as Wiebe and associates did in English (2002), since the agreement

TABLE 9. Agreement of annotations of news and blog articles.

Source	NTCIR		Blog	
Level	Sentence	Document	Sentence	Document
Agreements of two annotators	53.33%	41.18%	73.85%	64.71%
All agree	33.33%	17.65%	61.40%	41.18%

TABLE 10. Summary of the gold standard.

	Positive	Neutral	Negative	Nonopinionated	Total
Word	256	27	243	312	838
Sentence	48	3	93	432	576
Document	7	2	11	14	34

DOI: 10.1002/asi

decreases to an unacceptable degree when more annotators and more categories are involved.

The annotations are called *strongly inconsistent* if positive polarity and negative polarity are assigned to the same word by different annotators. In a total of 385 inconsistent answers, only 16 words are strongly inconsistent (4.16%). In contrast, annotations are highly inconsistent in weak opinion words of the form *positive/negative* vs. *neutral* (30) and *sentiment* vs. *nonsentiment* (339). It also shows that deciding the opinion degree of a word is challenging for human annotators.

The agreements at sentence level and document level are shown in Table 7 and Table 8, respectively. The results are quite similar to that at word level. Furthermore, agreements of annotations from news and blog articles are listed in Table 9 for comparison. Table 9 shows that tagging news articles achieves lower agreement rates than tagging Web blogs. We observe that blog articles use more daily vocabulary and are more easily understood by human annotators than news articles.

From the analyses of interannotator agreement, we find that the agreement drops fast when the number of annotators increases. It is less possible to have consistent annotations when more annotators are involved. Here we adopt voting to create a gold standard. The majority of the annotations of one instance are taken as a gold standard for the latter evaluation. If the annotations of an instance are all different, this instance is dropped. A total of three documents and 18 sentences are neglected. According to this criterion, Table 10 summarizes the statistics of the annotated testing data.

TABLE 11. Interpretation of kappa values.

Kappa value	Meaning
<0	Less than chance agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

TABLE 12. Kappa values of the word level agreement.

Two annotators	Kappa values	Annotator vs. gold standard	Kappa values
A vs. B	0.41	A vs. G	0.79
A vs. C	0.69	B vs. G	0.56
B vs. C	0.48	C vs. G	0.89
Average	0.53	Average	0.75

TABLE 13. Kappa values of the sentence level agreement.

Two annotators	Kappa values	Annotator vs. gold standard	Kappa values
A vs. B	0.17	A vs. G	0.49
A vs. C	0.44	B vs. G	0.59
B vs. C	0.41	C vs. G	0.83
Average	0.34	Average	0.64

TABLE 14. Kappa values of the document level agreement.

Two annotators	Kappa values	Annotator vs. gold standard	Kappa values
A vs. B	0.18	A vs. G	0.52
A vs. C	0.38	B vs. G	0.66
B vs. C	0.31	C vs. G	0.79
Average	0.29	Average	0.66

Kappa Value Analysis

We further assess the usability of the tagged corpus by kappa values. Kappa value gives a quantitative measure of the magnitude of interannotator agreement. Table 11 shows a commonly used scale of the kappa values. Tables 12–14 show the kappa values of the agreement at word, sentence, and document levels for pairwise checking between two annotators as well as an annotator and gold standard.

In Tables 12 to 14, the kappa values at sentence and document levels are smaller than those at word level. The annotations at all levels achieve substantial agreement on average with the gold standard i.e., between 0.61 and 0.80. The results ensure the usability of the evaluation corpus.

Evaluations of Opinion Extraction

The opinion related tasks involve much human perspective. To evaluate the proposed opinion extraction algorithms, the following sections show the performance of a human being, say, “ideal performance,” and then the real performance of the extraction algorithms.

Ideal Performance of Opinion Extraction

Before evaluating automatic opinion extraction algorithms, we have to know the ideal performance of human beings in the same task. Tables 15–17 list the evaluation of the three annotators with respect to the gold standard. Apparently, none of the annotators can assign 100% of the same answers as the gold standard, that is, the majority. These

TABLE 15. Annotators’ performance referring to gold standard at word level.

Annotators	Recall	Precision	f-Measure
A	94.29%	80.51%	86.86%
B	96.58%	88.87%	92.56%
C	52.28%	73.17%	60.99%
Average	81.05%	80.85%	80.14%

TABLE 16. Annotators’ performance referring to gold standard at sentence level

Annotators	Recall	Precision	f-Measure
A	94.44%	71.20%	81.19%
B	38.89%	74.67%	51.14%
C	90.97%	50.19%	64.69%
Average	74.77%	65.35%	65.67%

DOI: 10.1002/asi

TABLE 17. Annotators' performance referring to gold standard at document level.

Annotators	Recall	Precision	f-Measure
A	100%	71.43%	83.33%
B	50%	71.43%	58.82%
C	85%	65.38%	73.91%
Average	78.33%	69.41%	72.02%

results reveal an interesting observation. In the opinion extraction task, one annotator cannot know the opinion of the whole. The statistics tell us that the agreement between one annotator's opinions and the majority on average is around 80%. The other 20% is inconsistent because of annotators' own perspective. Such phenomena make opinion extraction different from other research topics in information retrieval and extraction.

System Performance of Opinion Extraction

The gold standard in Table 10 is used to evaluate the performance of opinion extraction at word, sentence, and document levels. The performance of the algorithm proposed in this paper is compared with two machine learning algorithms, i.e., Support Vector Machine (SVM) and the decision tree, at word level. See 5 (C5.0) system is employed to generate the decision tree. For the proposed extraction algorithm, qualified seeds are used for training (set A) and the gold standard is used for testing (set B). For machine learning algorithms, both qualified seeds and the gold standard are used for training and testing. The experimental results are reported in Tables 18–20.

As Tables 18–20 show, the proposed sentiment word mining algorithm achieves the best average precision, 61.06% of Verb and Noun, while SVM achieves 46.81% (outside test, the average of 45.23% and 48.39%), and C5 does even worse (precision 0% because of a small training set). Our algorithm outperforms SVM and the decision tree methods in sentiment word mining. The experiments indicate that the semantics within a word is not enough for a machine learning classifier. In other words, machine learning methods are not suitable for word level opinion extraction. In the past, Pang, Lee, and Vaithyanathan (2002) showed that machine learning methods are not good enough for opinion extraction at document level. Thus we may conclude that opinion extraction is beyond a classification problem. Compared to Table 15, our sentiment word miner achieves 85.44% of the performance of human annotators (i.e., 0.6847/0.8014).

TABLE 18. Performance of sentiment word mining.

	Nonnormalized			Normalized		
%	Verb	Noun	Average	Verb	Noun	Average
Precision	69.25	50.50	59.88	70.07	52.04	61.06
Recall	75.48	81.45	78.47	76.57	82.26	79.42
f-Measure	72.23	62.35	67.29	73.18	63.75	68.47

TABLE 19. Performance of SVM (precision).

		Testing	
SVM		A	B
Training	A	92.08%	45.23%
	B	48.39%	56.87%

TABLE 20. Performance of decision tree (precision).

		Testing	
C5		A	B
Training	A	83.60%	36.50%
	B	0%	41.50%

Table 21 and Table 22 further show the results of opinion extraction at sentence and document levels. Results, including news and blog articles, indicate that the precision rates are low at both sentence and document levels. This is because the algorithm so far only considers opinions and not relevance. Many sentences that are nonrelevant to the topic "animal cloning" are included for opinion judgment. The nonrelevant rate is 50% and 53% for news articles and Web blog articles, respectively.

The quantity of seeds also influences the performance of opinion extraction. Figure 5 shows their relationship. The more seeds are used, the better performance is achieved. However, Figure 5 further indicates that the improvement of performance converges when the quantity of seeds reaches around 8,000. Recall that 10,542 qualified seeds are used in this paper. Thus, we can neglect the issue of the lack of seeds.

As mentioned, extracting opinions only is not sufficient for opinion summarization. The focus of opinions should also be considered. In the following opinion summarization section, an algorithm for detecting relevant sentences is introduced and applied when extracting sentences for opinion summarizations. The experimental results of opinion extraction considering relevance are also listed.

TABLE 21. Opinion extraction at sentence level.

Source	NTCIR	BLOG
Precision	34.07%	11.41%
Recall	68.13%	56.60%
f-Measure	45.42%	18.99%

TABLE 22. Opinion extraction at document level.

Source	NTCIR	BLOG
Precision	40.00%	27.78%
Recall	54.55%	55.56%
f-Measure	46.16%	37.04%

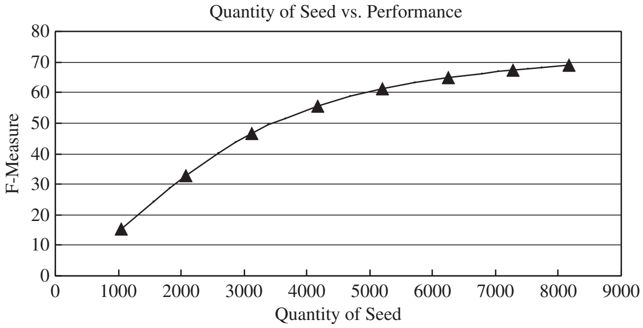


FIG. 5. The relationship between the performances of opinion extraction and the quantity of seeds.

Opinion Summarization

Traditional summarization algorithms (Chen, Kuo, Huang, Lin, & Wung, 2003) focus on similarity/dissimilarity computation among sentences and remove the redundant information. Unlike in the traditional algorithms, two factors—the sentiment degree and the correlated events—play the major roles in opinion summarization. The repeated opinions of the same polarity cannot be dropped because they strengthen the sentiment degree. In contrast, the redundant reasons why they hold this position can be removed when generating opinion summaries. And needless to say, opinions must be detected first for the opinion summarization. All these issues make opinion summarization challenging.

Opinion summarization aims to produce a cross-document opinion summary. For this purpose, it is necessary to know which sentences express opinions and determine whether they focus on a designated topic. An algorithm that decides the relevance degree and the sentiment degree is proposed in this section. To put emphasis on the opinion factor, the visualization of opinion summaries is different from the traditional ones. A text-based summary categorized by opinion polarities is illustrated in this section. Then, a graph-based summary along time series is illustrated by an opinion tracking system introduced in Opinion Tracking System.

Summarization Algorithm

Both topical information and opinion information are considered in the algorithm of generating opinion summaries. Here topical words are selected to represent the concept of one topic. The relevance of every sentence is decided according to the selected topical words. Then the algorithm of opinion extraction at sentence level is applied to extract suitable sentences for summarization.

Selection of topical words. Choosing representative words that can exactly present the main concepts of a relevant document set is the main work of relevant sentence retrieval. A term is considered to be representative if it appears frequently across documents or appears frequently in each document (Fukumoto & Suzuki, 2000). Such terms form the major topic of the relevant document set. The major topic is chosen as follows: We assign weights to each word at both document level and paragraph level. In the following formulas, ***W*** denotes weights, ***S*** denotes document level, and ***P*** denotes paragraph level. ***TF*** is term frequency, and ***N*** is document count in formula 10 and paragraph count in formula 11. In the subscripts, symbol *i* is the document index, symbol *j* is the paragraph index, and symbol *t* is the word index. Formulas 10 and 11 compute ***TF*IDF*** scores of term *t* in document *i* and paragraph *j,* respectively. Formulas 12 and 13 denote the frequency at which term *t* appears across documents and paragraphs. Formulas 14 and 15 denote the frequency at which term *t* appears in each document and in each paragraph.

$$W_{s_it} = TF_{s_it} \times \log \frac{N}{N_{s_i}} \tag{10}$$

$$W_{P_jt} = TF_{P_jt} \times \log \frac{N}{N_{P_j}} \tag{11}$$

$$Disp_{St} = \sqrt{\frac{\sum_{i=1}^{m} (W_{S_it} - mean)^2}{m} \times TH} \tag{12}$$

$$Dev_{s_it} = \frac{W_{s_it} - mean}{Disp_{St}} \tag{13}$$

$$Disp_{Pt} = \sqrt{\frac{\sum_{j=1}^{n} (W_{P_jt} - mean)^2}{n}} \tag{14}$$

$$Dev_{P_jt} = \frac{W_{P_jt} - mean}{Disp_{Pt}} \tag{15}$$

A term is thought to be representative if it satisfies either formula 16 or formula 17. Terms satisfying formula (16) tend to appear in few paragraphs of many documents, while terms satisfying formula (17) appear in many paragraphs of few documents. The score of a term, defined as the absolute value of Dev_{P_jt} minus Dev_{S_it}, measures how significant it is to represent the main concepts of a relevant document set.

$$Disp_{St} \leq Disp_{Pt} \;\; \exists S_i, \forall P_j \in S_i Dev_{S_it} \leq Dev_{P_jt} \tag{16}$$

$$Disp_{St} > Disp_{Pt} \;\; \exists S_i, \forall P_j \in S_i Dev_{S_it} > Dev_{P_jt} \tag{17}$$

Comparing our findings with those of Fukumoto and Suzuki (2000), we modify the scoring function at paragraph level. All documents in the same corpus are concatenated into a big one; i.e., the original boundaries between documents are ignored, so that words that repeat frequently among paragraphs are chosen. We also use threshold ***TH*** to control the number of representative terms in a relevant corpus. The larger the threshold ***TH*** is, the more terms are included. The value of this parameter is trained in the experiments. The selected topical words are then used in relevant sentence retrieval. The corpus of the novelty track of TREC is used for evaluation (Soboroff & Harman, 2003).

The method of considering only topical words is competitive with that of the top three results over 55 submissions. The performance of extracting relevant sentences is comparably good (Ku, Lee, Wu, & Chen, 2005a).

Generation of opinion summaries. In the opinion summarization algorithm, sentences that are relevant to the major topic and express opinions are extracted. There are two clues for extraction: concept keywords and sentiment words. The former determine the relevance of a sentence to the topic, and the latter identify the degree of its opinions. In our experiments, concept keywords are from a predefined field (NTCIR) or automatically extracted from documents on the same topic, that is, set 2 of TREC, topic ZH037 of NTCIR, and "animal cloning" for blog articles (BLOG).

In opinion summarization, we consider the opinion tendencies of all relevant sentences. The relevance degree and the opinion degree of a sentence do not interfere with each other. Selection of Topical Words mentioned that retrieving relevant sentences according to the selected topical words has a satisfactory performance. Hence instead of calculating relevant scores, sentences are considered relevant if they contain topical words. Sentences expressing opinions are extracted from the topical sentence set and their tendencies are determined. Compared to the sentence-level opinion extraction algorithm in Sentence Level, detecting the topical sentences is the first step here. The overall procedure is as follows.

Algorithm: *Topical Opinion Sentence Extraction*

```
1. For every topical sentence $s$
2.     For every sentiment word $w$ in $s$
3.         If a negation operator appears nearby, reverse the
           sentiment tendency. Every sentiment word $w$
           contributes its sentiment score to $s$.
4. Decide the opinion tendency of $s$ by the functional
     composition of sentiment words $w$s, i.e., Formula (7).
```

Experiments and Illustration of Summaries

Sentiment words are used to decide opinion polarities. If the total score of sentiment words is positive/negative, the sentence is positive/negative-oriented. We also consider opinion operators, e.g., *say*, *present*, *show*, *suggest.* At least one opinion holder exists when opinion operators appear in one sentence. If a sentence contains such an opinion operator that follows a named entity and this sentence is of 0 opinion score, it is regarded as a neutral opinion.

As mentioned, major topic detection is required for opinion summarization. Table 23 shows the results of considering relevance together with sentiments. NTCIR corpus, with TREC style, contains concept words for each topic. These words are taken as the topical words for the opinion extraction. Sentences containing at least one concept word are considered topical, that is, relevant to the major topic.

Obviously, the results are much better than those in Tables 21 and 22. However, in the real world applications, the major topics, concepts, or topical words may not be available. For Web blog articles or news stories, words that represent the major topic must be selected automatically. The algorithm for choosing representative words is adopted and opinion sentences are extracted to show the performance again. Table 24 shows the experimental results.

When it is compared to Table 21, the precision increases after applying major topic detection algorithm. It concludes that the relevant sentence selection is important in the opinion summarization. If the sentence is not relevant to the topic we are interested in, the extracted opinions are not useful.

A total of 29.67% and 72.43% of nonrelevant sentences are filtered out for news and Web blog articles, respectively. The performance of filtering nonrelevant sentences in blog articles is better than that in news articles. It is also consistent with the higher agreement rate of annotations in blog articles. A total of 15 topical words are extracted automatically from blog articles while more, 73 topical words, are extracted from news articles. These all indicate that the content of news articles diverges more than that of blog articles when describing the same issue. However, the judgment of sentiment polarity of blog articles is not simpler (in Table 24 precision of 38.06% vs. 23.48%).

The topical information and the sentiment degree of each sentence are employed to generate opinion summaries. We distinguish between positive and negative documents. A document is positive if it consists of more positive-topical sentences than negative-topical ones; and vice versa. Among positive and negative documents, two types of opinion summaries are proposed: brief and detailed opinion summaries. For the brief summary, we pick up the document with the largest number of positive or negative topical sentences and use its headline to represent the overall summary of positive-topical or negative-topical sentences. For the detailed summary, we list positive-topical and negative-topical sentences with higher sentiment degree. Examples of brief and detailed summaries are shown in Tables 25 and 26, respectively.

TABLE 23. Extraction considering concept words.

Source	NTCIR	
Level	Sentence	Document
Precision	57.80%	76.56%
Recall	67.23%	72.30%
f-Measure	62.16%	74.37%

TABLE 24. Extraction considering extracted topical words.

Source	NTCIR	Blog
Precision	38.06%	23.48%
Recall	64.84%	50.94%
f-measure	47.97%	32.58%

DOI: 10.1002/asi

Comparing Opinion Summaries of News and Blogs

Compared to the opinion summary of TREC set 2 (shown in Table 26), Tables 27 and 28 list the opinion summaries of NTCIR and blog articles using the same topic, "animal cloning".

News and blog articles are two main sources for opinions on the Web. Different sources of articles enrich the content. Generally speaking, news documents are more objective while blog articles are usually more subjective. Besides, the opinion holders from two sources are of different social classes. The opinions extracted from news are mostly from famous people, whereas the sources of opinions expressed in blogs may not be identified by name. Listing opinions of different sources in parallel provides various views of the same public issue.

The proposed algorithm of opinion summarization is language independent. With this method, opinions of different countries are visible, as Tables 26–28 have shown. Moreover, this is a prototype of the cross-lingual opinion summarization.

TABLE 25. Brief opinion summary of "clone Dolly sheep" in TREC corpus.

Positive	Chinese Scientists Suggest Proper Legislation for Clone Technology
Negative	UK Government Stops Funding for Sheep Cloning Team

TABLE 26. Detailed opinion summary of "clone Dolly sheep" in TREC corpus.

Positive	Ahmad Rejai Al-Jundi, Assistant Secretary General of the Islamic Organization, declared earlier that the seminar would be aimed at shedding light on medical and legal aspects of the internationally controversial issue and seeking to take a stand on it.
Negative	Dolly the cloned sheep is only 3, but her genes are already showing signs of wear and she may be susceptible to premature aging and disease—all because she was copied from a 6-year-old animal, Scottish researchers say.

Opinion Tracking System

Although opinion summaries can be generated by using algorithms of opinion extraction and opinion summarization, they may be distributed discretely when the quantity of relevant documents is large. As for events, we are more concerned of how opinions change over time. An opinion tracking system aims to tell how people change their opinions as time goes by.

System Framework

Figure 6 shows the framework of an opinion tracking system. Documents relevant to the opinion request are retrieved at the first step. Documents then are arranged in a chronological order by the temporal processing module for opinion extraction. With this temporal information, the summarization module identifies which events are correlated with the designated opinions.

The core of this system is the opinion extraction module. Using sentiment dictionaries, sentiment scores of characters are calculated. The sentiment miner extracts evidence at word level according to the sentiment scores of characters. Then the opinion extraction module extracts opinions at both sentence and document levels. Algorithms in Opinion Extraction Algorithms are employed. Using sentiment scores along with temporal information, the opinion summarization module generates a statistical report, summarizes the opinions, and places the correlated major events along a timeline.

TABLE 27. Detailed opinion summary of "animal cloning" in NTCIR corpus.

Positive	上述建議來自四名科學家所組成的專家小組, 該小組於一月應英國政府之邀成立, 就複製所衍生的法律與倫理問題提出相關建議。 (The above suggestion came from a group of four scientists. The group was formed under the request of the British government. The group was to provide advice on laws and theories concerning cloning.)
Negative	在複製羊成功的消息宣布之後, 美國總統柯林頓及「生物倫理顧問委員會」斥複製人不道德 柯林頓禁止使用聯邦經費從事複製人類的實驗, 並要求民間自我克制不作這種研究 。 (After the announcement of the success in sheep cloning, U.S. President Clinton and National Bioethics Advisory Commission reproved human cloning as immoral. Clinton forbade using federal funds for human cloning experiments and asked the general public to refrain from doing such research.)

TABLE 28. Detailed opinion summary of "animal cloning" in BLOG corpus.

Positive	而複製技術如果成熟, 它將會是一種強大有用的工具, 任何工具都可能被善用或誤用, 評價一個工具不能只拿它被誤用的情境去批評它, 因而禁制了它被善用的原始目的與機會, 妥善的立法規範管理似乎才是較理性的作為。 (When the cloning technology reaches maturity, it will become a powerful and useful tool. Any tool can be used for a good cause or misused, so we should not blatantly criticize and dismiss a tool for its possible abuses and deprive it of its opportunity to be used in a good way. Instead, we should come up with suitable regulations for using the tool.)
Negative	有人反對複製人, 因為違反了上帝的旨意。 (Some people are against cloning human beings, because it conflicts with teachings of God.)

System Illustration

A certain quantity of relevant articles is necessary for tracking opinions. Because the number of articles relevant to “animal cloning” in Table 4 is not large enough for demonstration, we take the year 2000 presidential election in Taiwan as an example. Four persons’ names are used as queries to the information retrieval (IR) system and opinions about them in March 2000 are shown in Figure 7.

Persons A, B, and C were candidates and D was the president at that time. Person A was the president-elect on Election Day. The trend fits the opinions in this period and the opinion summaries indicate events correlated with these opinions. This tracking system also tracks opinions according to different requests (topics) and different information sources, including news agencies and the Web. Opinion trends of the same topic but from different expressers can also be compared, in a similar form as in Figure 7. This information is very useful for the government, institutes, companies, and the concerned public.

Conclusion and Future Work

This paper proposes algorithms for opinion extraction, summarization, and tracking. Different materials in different languages are experimented with and compared. Two major document types, news and blog articles, are chosen as experiment materials in this paper. The natures of news and blog articles are quite different. News articles, compared to blog articles, have a larger vocabulary. The quantity of the extracted topical words indicates the same thing. On one hand, that makes the relevant sentence retrieval harder. On the other hand, a larger vocabulary helps when deciding sentiment polarities.

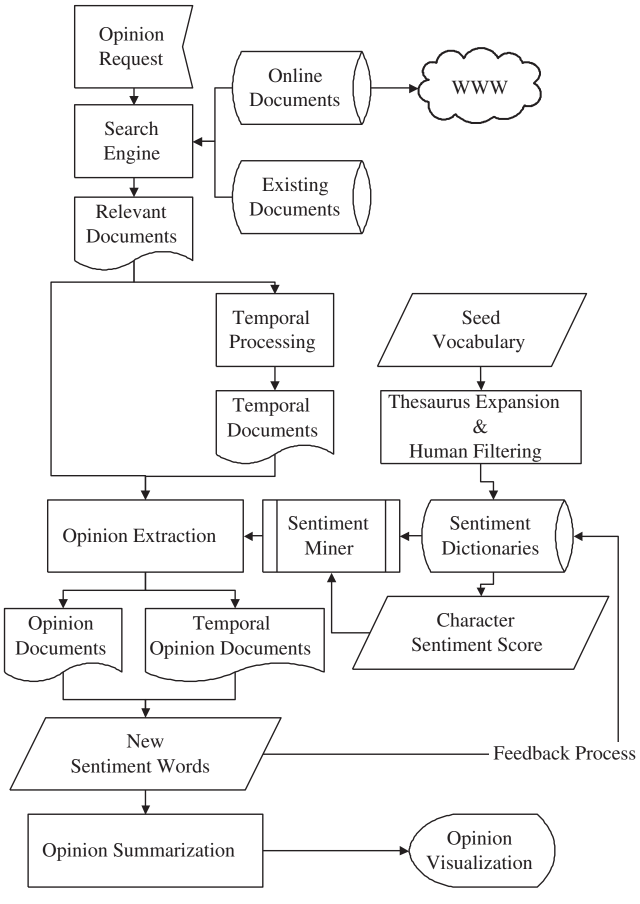


FIG. 6. Framework of an opinion tracking system.

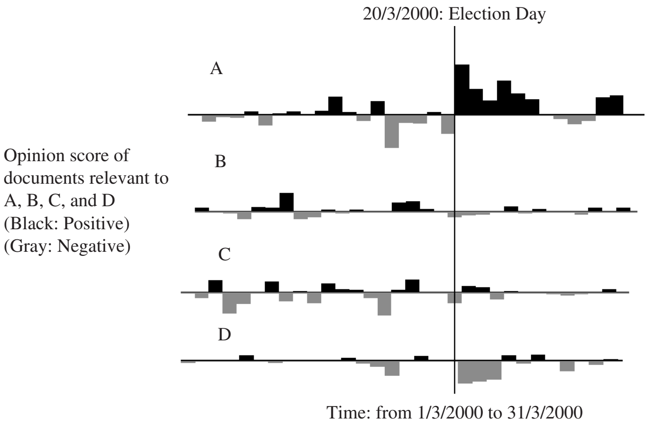


FIG. 7. Opinions about four persons in a presidential election.

There are three different granularities for opinion mining: word, sentence, and document. Positive and negative sentiment words and their weights are mined on the basis of Chinese word structures. The f-measure is 73.18% and 63.75% for verbs and nouns, respectively. Experimental results also indicate that machine learning methods are not suitable for word-level sentiment mining. Utilizing the mined sentiment words together with topical words, we achieve f-measure 62.16% at the sentence level and 74.37% at the document level. With topical words involved, the performance of opinion extraction is enhanced.

An opinion tracking system provides not only text-based and graph-based opinion summaries, but also the trend of opinions from different information sources. Opinion summaries show the reasons for different stands people take on public issues. Comparisons among opinion summaries in different languages and from different sources help to identify the attitudes of people in different groups.

Opinion holders are considered in this work. Experts or government officers have more influence when expressing opinions. However, the ways opinion expressers influence the sentiment degree has not yet been explored. Identifying opinion holders is very important for analyzing opinions. Properly deciding the power of opinion holders not only indicates sentiment degree reliably, but also answers opinion questions. Moreover, the relations between holders and their opinions are the key to solving multiperspective problems related to opinions.

Because the major goal of this paper is to study opinion mining under different factors, we made experiments using relevant document sets and ignored the issues of integrating document retrieval. In real world applications, the performance of the front-end IR system affects the performance of the back-end opinion mining system. Which IR system, i.e., highly precise, good recall, or balance of both, is suitable for an opinion mining system will be investigated in the future.

Acknowledgments

Research for this paper was partially supported by National Science Council, Taiwan, under contract NSC95-2221-E-002-265-MY3 and Excellent Research Projects of National Taiwan University, under the contract 95R0062-AE00-02.

References

Bai, X., Padman, R., & Airoldi, E. (2005). On learning parsimonious models for extracting consumer opinions. In Proceedings of the 38th Annual Hawaii International Conference on System Sciences (track 3, vol. 03, p. 75.2). Los Alamitos: IEEE Press.

Cardie, C., Wiebe, J., Wilson, T., & Litman, D. (2003). Combining low-level and summary representations of opinions for multi-perspective question answering. In Proceedings of 2003 AAAI Spring Symposium on New Directions in Question Answering (pp. 20–27). Menlo Park: AAAI Press.

Chen, H.-H., Kuo, J.-J., Huang, S.-J., Lin, C.-J., & Wung, H.-C. (2003). A summarization system for Chinese news from multiple sources. Journal of American Society for Information Science and Technology, 54(13), 1224–1236.

Chen, K.-H., & Chen, H.-H. (2001). Cross-language Chinese text retrieval in NTCIR workshop—towards cross-language multilingual text retrieval. ACM SIGIR Forum, 35(2), 12–19.

Dave, K., Lawrence, S., & Pennock, D.M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th International World Wide Web Conference (pp. 519–528). New York, NY: ACM Press.

Fukumoto, F., & Suzuki, Y. (2000). Event tracking based on domain dependency. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 57–64). New York, NY: ACM Press.

Hiroshi, K., Tetsuya, N., & Hideo, W. (2004). Deeper sentiment analysis using machine translation technology. In Proceedings of the 20th International Conference on Computational Linguistics (pp. 494–500). Morristown, NJ: Association for Computational Linguistics.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 168–177). New York, NY: ACM Press.

Kim, S.-M., & Hovy, E. (2004). Determining the sentiment of opinions. In Proceedings of the 20th International Conference on Computational Linguistics (pp. 1367–1373). Morristown, NJ: Association for Computational Linguistics.

Kishida, K., Chen, K.-H., Lee, S., Chen, H.-H., Kando, N., Kuriyama, K., et al. (2004). Cross-lingual information retrieval task at the NTCIR Workshop 3. ACM SIGIR Forum, 38(1), 17–20.

Ku, L.-W., Lee, L.-Y., Wu, T.-H., & Chen., H.-H. (2005a). Major topic detection and its application to opinion summarization. In Proceedings of 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 627–628). New York, NY: ACM Press.

Ku, L.-W., Wu, T.-H., Lee, L.-Y., & Chen., H.-H. (2005b). Construction of an evaluation corpus for opinion extraction. Proceedings of the Fifth NTCIR Workshop Meeting (pp. 513–520). Tokyo: National Institute of Informatics.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion Observer: Analyzing and comparing opinions on the Web. In Proceedings of the 14th International World Wide Web Conference (pp. 342–351). New York, NY: ACM Press.

Mei, J., Zhu, Y., Gao, Y., & Yin, H. (1982). tong2yi4ci2ci2lin2. Shanghai: Shanghai Dictionary Press.

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 (pp. 79–86). Morristown, NJ: Association for Computational Linguistics.

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (pp. 105–112). Morristown, NJ: Association for Computational Linguistics.

Riloff, E., Wiebe, J., & Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. In Proceedings of the Seventh Conference on Natural Language Learning (pp. 25–32). Edmonton, Canada: Morgan Kaufman Publishers.

Soboroff, I., & Harman, D. (2003). Overview of the TREC 2003 novelty track. In Proceedings of the 12th Text REtrieval Conference, National Institute of Standards and Technology (pp. 38–53). Washington, DC: U.S. Government Printing office.

Takamura, H., Inui, T., & Okumura, M. (2005). Extracting semantic orientations of words using spin model. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (pp. 133–140). Morristown, NJ: Association for Computational Linguistics.

Wiebe, J., Breck, E., Buckly, C., Cardie, C., Davis, P., Fraser, B., et al. (2002). NRRC summer workshop on multi-perspective question answering, final report. In ARDA NRRC Summer 2002 Workshop. (1–64).

Wiebe, J., Wilson, T., & Bell, M. (2001). Identify collocations for recognizing opinions. In Proceedings of ACL/EACL2001 Workshop on Collocation. Computational Extraction, Analysis, and Exploitation. (pp. 24–31). Morristown, NJ: Association for Computational Linguistics.

DOI: 10.1002/asi