Aslib Journal of Information Management

Sentiment search: an emerging trend on social media monitoring systems

Tung Thanh Nguyen, Tho Thanh Quan, Tuoi Thi Phan,

Sentiment search: an emerging trend on social media monitoring systems

Tung Thanh Nguyen, Tho Thanh Quan and Tuoi Thi Phan
Department of Computer Science, University of Technology, Ho Chi Minh, Vietnam

Received 15 December 2013
Revised 7 July 2014
20 July 2014
Accepted 23 July 2014

Abstract

Purpose – The purpose of this paper is to discuss sentiment search, which not only retrieves data related to submitted keywords but also identifies sentiment opinion implied in the retrieved data and the subject targeted by this opinion.

Design/methodology/approach – The authors propose a retrieval framework known as Cross-Domain Sentiment Search (CSS), which combines the usage of domain ontologies with specific linguistic rules to handle sentiment terms in textual data. The CSS framework also supports incrementally enriching domain ontologies when applied in new domains.

Findings – The authors found that domain ontologies are extremely helpful when CSS is applied in specific domains. In the meantime, the embedded linguistic rules make CSS achieve better performance as compared to data mining techniques.

Research limitations/implications – The approach has been initially applied in a real social monitoring system of a professional IT company. Thus, it is proved to be able to handle real data acquired from social media channels such as electronic newspapers or social networks.

Originality/value – The authors have placed aspect-based sentiment analysis in the context of semantic search and introduced the CSS framework for the whole sentiment search process. The formal definitions of Sentiment Ontology and aspect-based sentiment analysis are also presented. This distinguishes the work from other related works.

Keywords Semantic search, Sentiment Ontology, Aspect-based sentiment analysis, Incremental Ontology update, Lightweight NLP, Sentiment search

Paper type Research paper

Introduction

Semantic search

There is a huge demand for retrieval of meaningful information from very large data sets in both the research and industrial communities. For instance, in a request recently posted in NineSigma, a portal for exchanging ideas between the research and industrial communities, we learnt that a multi-billion dollar industrial plant facility company is seeking a software algorithm that uses a natural language query to retrieve matching results from large-scale time-series data sets created from measurements taken at industrial plant facilities[1]. From research point of view, this is a problem of semantic search where one needs to infer semantics from both queries and stored data in order to produce appropriate retrieval results.

There are various research works reported on the track of semantic search, of which we give a summary and analysis in the next section. However, it can be easily observed

This work is supported by the research project B0212-20-02TD funded by the Vietnam National University – Ho Chi Minh City. The authors are also grateful to YouNet Media for supporting real data sets for the experiments.

Aslib Journal of Information Management
Vol. 66 No. 5, 2014
pp. 553-580

2050-3806
DOI 10.1108/AJIM-12-2013-0141

that the concept of semantics varies depending on the search purpose. In a general search engine like Google[2], the semantics are associated with the meaning of the queried keywords. Meanwhile, in WolframAlpha's system[3], the semantics are initially inferred from the given mathematical expressions, and then evolved into the so-called computational knowledge, i.e. knowledge that is computable on computer systems. In a narrower sense, the semantics in a mobile pathfinder application is limited to certain location-based concepts such as city, street, nearest restaurant, shortest path, etc.

Sentiment search

In this paper, we focus on a particular case of semantic search, known as sentiment search, in which the retrieval system is capable of not only returning objects (like keywords, images, etc.) relevant to a query, but also inferring the attitudes of authors/writers when mentioning the objects in the search results. Those attitudes can be positive (e.g. "I enjoy this smartphone very much"), negative (e.g. "This smartphone makes me really disappointed") or neutral ("There is an advertisement of this smartphone on TV").

Nowadays, sentiment search has been emerging as a trend on social monitoring systems, like Social Mention[4] or Jive[5]. In such systems, users can search for certain topics or objects mentioned on popular social media channels like electronic newspapers or social networks. Thus, those systems are very useful for marketers or brand managers, who always want to observe public opinions of certain brands or products, or even features of products. Obviously, in this case, sentiment information inferred from the corresponding mentions is very crucial.

However, as compared to other directions of semantic search, the research on sentiment search is currently suffering from certain drawbacks. Typically, there are three major approaches which can be fused together for semantic search. They include: semantic-based approach (like using domain ontology); lightweight Natural Language Processing (NLP) techniques; and data mining techniques using statistics/information theory. For sentiment search, empirical results pose certain difficulties for these three approaches (Socher *et al.*, 2013; Liu, 2012; Glorot *et al.*, 2011). Regarding using domain ontologies, each domain introduces different sentiment concepts which could even be contradictory to each other. For example, the concept of hot may imply positive sense for a smartphone product ("This smartphone is currently very hot on the market") but indicating negative sense for a food ("This food is too hot for children"). In the meantime, the linguistic phrases/structures used to imply sentiment sense are also more complex due to different contexts of discussion. For example, the statement "next time I will try other foods" implies a negative sense for the food currently discussed even though no sentiment phrase is actually given. This reason also introduces a considerable obstacle if one wants to give a cross-domain data mining model for sentiment classification.

Contributions

In this paper, we make the following research contributions:

(1) We present a common framework for semantic search. We show that the proposed framework captures all major aspects and directions of research on semantic search reported so far.

(2) We present a survey of the current studies of sentiment analysis.

(3) We formally define Sentiment Ontology and subsequently use it for the formal definition of aspect-based sentiment analysis problem. To the best of our knowledge, this is the first time those concepts are formally defined.

(4) We evolve the common framework of semantic search into a framework known as Cross-Domain Sentiment Search (CSS), which can overcome the current difficulties of sentiment search. The major features of CSS include the use of domain ontologies, so-called Industry Ontologies, and the definitions of specific linguistic rules, so-called Sentiment Phrasing Rules, to handle sentiment phrases.

(5) We introduce an incremental approach that can be used to automatically adapt our proposed sentiment search to new domains. Thus, our technique can potentially solve the prominent cross-domain problem of sentiment search.

Common framework for semantic search

In Figure 1, we present a common framework which captures major steps of semantic search process. By a subsequent literature survey, we will show that various research directions in this area adopt this framework, partially or completely. The framework consists of the following major components.

Domain knowledge resources

This component contains structured resources that capture major concepts and their relations in the domain on which the semantic search techniques are to be applied. Most research works use domain ontologies (Gruber, 1993) for this purpose. The domain ontologies are mostly constructed manually, the main purpose being a more accurate semantic processing of the documents. WordNet (Miller *et al.*, 1990) also serves as a Linguistic Ontology in some works.

Semantic feature extraction

This component extracts features from documents for indexing and further processing. A feature is a piece of information that is useful for inferring the semantics of the documents. In approaches that employed ontologies, the semantic features to be extracted are mostly ontological concepts and relations. In other approaches, the extracted features are important factors of the documents such as titles or weighted important terms.

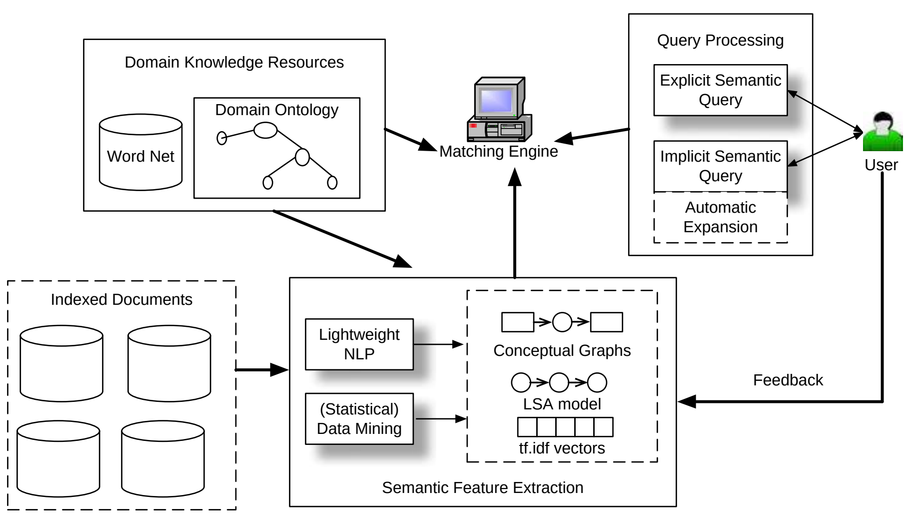

Figure 1.
The common framework of semantic search techniques

Typically, there are two major techniques used for feature extraction in semantic search as follows:

- *Lightweight NLP*: NLP techniques are always favoured for handling textual data, especially for ontology-based approaches. However, due to the extremely high computational cost of conventional NLP techniques, lightweight methods such as shallow parsing (Abney, 1991) are traditionally adopted. The main objectives of those lightweight methods are to identify the constituents (noun groups, verbs, verb groups, etc.), but not specifying precisely their internal structure.
- *Data mining techniques and/or statistical methods*: traditional data mining methods which may involve statistical factors such as tf.idf weighting (Salton and Buckley, 1988) or Bayesian inferrence (Hazewinkel, 2001) are quite useful for classical search engines. For semantic search, they are often used as supporting methods, mostly combined with ontology domain to capture concept candidates of the domain, as subsequently illustrated in our summarized survey.

After extraction, semantic features are then organized as a conceptual structure used for indexing and retrieval. Common conceptual structures used for semantic search include conceptual graphs (CG) (Sowa, 1976), or vectorized semantic structures like Latent Semantic Analysis (LSA) model (Dumais, 2004) or tf.idf weighting vectors. It is noted that in the modern information retrieval systems (e.g. the famous Google), the parameters used in data mining techniques and statistical methods are not remained fixed once the system actually executes with actual data. Based on the relevance feedback from users, those parameters can be adjusted accordingly to get better accuracy.

Query processing

The major purpose of this method is to infer the semantic implication of the queries submitted by users. Generally, there are two major approaches for this component:

- *Explicit implication*: the semantics of the queries are expressed explicitly via well-structured query languages or a guided interface.
- *Implicit implication*: the queries submitted by users are still keywords-based. Then, the keywords are expanded to be compatible with the conceptual structures adopted by the semantic feature extraction components.

A summarized survey

Table I presents our survey of some works in semantic search. Limited by the scope and space of the paper, we do not discuss these works in detail. Instead, we only present their major techniques, with respect to the discussed common semantic search framework.

Sentiment analysis

Sentiment analysis (Nasukawa and Yi, 2003) or opinion mining (Dave *et al.*, 2003) is the task that aims to infer the sentiment orientation in a document (Padmaja and Fatima, 2013). For semantic search, this task plays the roles of semantic feature extraction, with respect to our presented common framework. The semantic features to be extracted in this case are the sentiment opinions. There are three levels of sentiment analysis as follows:

(1) document-based level;

(2) sentence-based level; and

(3) aspect-based level.

Work	Domain knowledge	Semantic features	Query processing
TAP (Guha *et al.*, 2003)	Domain ontology	Data aggregation	Keywords
XSEarch (Cohen *et al.*, 2003)	XML-based data sets		XML-based query language
SemSearch (Lei *et al.*, 2006)	Domain Ontology	Lightweight NLP	Keyword expansion
Li *et al.* (2002)	Linguistic Ontology	Conceptual graphs (CG)	CG matching
Paolucci *et al.* (2005)	Web services in OWL-S	UDDI registration	OWL-S/UDDI matchmaker
Cimiano *et al.* (2007)	Domain ontology		DL conjunctive queries
SPARK (Wang *et al.*, 2007)	Domain ontology	Conceptual graph	Formal logic queries
AVATAR (Kandogan *et al.*, 2006)	Structure datastore	Lightweight NLP	Keyword expansion
MultimediaN E-Culture (Schreiber *et al.*, 2008)	Cultural-heritage ontology	Conceptual graph	Graph matching
Q2Semantic (Liu *et al.*, 2008)	Domain ontology	Lightweight NLP	Keyword expansion
Ontogator (Hyvönen *et al.*, 2004)	Domain ontology	Lightweight NLP	Keyword expansion
K-search (Bhagdev *et al.*, 2008)	Domain ontology	Lightweight NLP	Keyword expansion
OntoWiki (Auer *et al.*, 2006)	Semantic Wiki	Collaborative approach	
XML Fragments (Duboue *et al.*, 2006)		Lightweight NLP	XML-based language
Fernandez *et al.* (2008)	QA Ontology	Lightweight NLP	Simple natural language
Semantic Wiki Search (Haase *et al.*, 2009)	Semantic Wiki	Wiki structure	structured query
Query-Based Faceted Search (QFS) (Ferré and Hermann, 2011)	Domain ontology	Faceted features	LISQL query language
Remote Associates Test (Smith *et al.*, 2013)		Latent semantic analysis	Multiply-constrained queries
GoWeb (Dietze and Schroeder, 2009)	Domain ontology	Text mining	Keywords
Metacat (Berkley, 2009)	Ecological ontology	Meta-data features	Keyword expansion
Falconer (Qu *et al.*, 2010)	Semantic annotation	Collaborative approach	SPARQL query language
Bruza *et al.* (2012)	Electronic medical records	Concept discovery	Concept matching
Embley *et al.* (2011)	Multilingual ontologies	Ontological mapping	

Table I. A survey of semantic search

Document-based sentiment analysis

Data mining/machine learning techniques are commonly used to infer sentiment opinion of a whole document. Those techniques always require training sets, which involve the following features:

- *Lexicon*: it involves both unigram and *n*-gram analysis.
- *Occurrence Frequency*: it typically involves tf.idf metric. In addition, some other works introduced BM25 idf, PMI and LSA (Kim *et al.*, 2009) and Delta tf.idf (Paltoglou and Thelwall, 2010).
- Some other features such as part-of-speech, location of words, sentiment terms such as "good," "wonderful," "bad" are also additionally used to support sentiment analysis.

The popular techniques used in the literature include Naive Bayes (Melville *et al.*, 2009; Xia *et al.*, 2011; Zhang *et al.*, 2011; Tan and Zhang, 2008; Ye *et al.*, 2009); support vector machine (SVM) (Xia *et al.*, 2011; Zhang *et al.*, 2011; Tan and Zhang, 2008; Prabowo and Thelwall, 2009; Xu *et al.*, 2011); maximal entropy (Go *et al.*, 2009; Gindl and Liegl, 2008; Shimada and Endo, 2008); and *n*-gram model (Ye *et al.*, 2009). Experimental results from these works suggest that SVM generally achieves the best performance for document-based sentiment analysis.

Sentence-based sentiment analysis

Data mining/machine learning techniques achieved good performance for document-based sentiment analysis, but generally failed to handle sentence-based sentiment due to lack of linguistic processing (Socher *et al.*, 2013). To handle this problem, Yu and Hatzivassiloglou (2003) suggested selection of a set of adjectives as seeds to infer sentiment implication of a sentence. This set of seeds can be expanded using distributional similarity (Lin, 1998). McDonald *et al.* (2007) proposed to use Conditional Random Field method, which was later extended as a set of labeled sentences (Täckström and McDonald, 2011).

Narayanan *et al.* (2009) argued that each kind of level should be treated differently when performing sentiment analysis. Not only that, the discourse information would also need to be taken into account to analyze multiple linked sentences. The common approach in handling this problem is to use lightweight NLP techniques such as shallow parsing to perform pattern mining for compound sentences (Asher *et al.*, 2008; Somasundaran *et al.*, 2008; Zhou *et al.*, 2011). Zirn *et al.* (2011) also proposed using Markov logic network to handle this problem.

Recently, deep learning (Hinton, 1990; Bottou, 2011) has been emerging due to its capabilities of producing data representations in several levels in a hierarchical manner. Thus, one can infer the meaning of a long phrase combined from shorter terms more precisely. In Glorot *et al.* (2011), deep learning has been proposed for domain adaptation in sentiment analysis. Especially, the Recursive Neural Tensor Network (Socher *et al.*, 2013) is claimed to achieve accuracy up to 84 percent, and is considered the best so far.

Aspect-based sentiment analysis

In document-based and sentence-based sentiment analysis, it is implicitly assumed that the analyzed document or sentence only discusses a single object. Therefore, the sentiment analysis process only focusses on inferring the sentiment opinion of the document/sentence, without identifying the object targeted by the opinion (Liu, 2012). Clearly these approaches would not be practically suitable when applied on real-life applications. For example, in a document, the author may at the same time give good comments on some features and bad comments on other features of a product. In this case, the sentiment analysis of this document may be neutral, but it is always desired to distinguish positive-rated features with negative-rated ones.

Motivated by this observation, Hu and Liu (2004), proposed to perform sentiment analysis at aspect-level. To be more precise, apart from rating positive/negative sense of a mention, the objects targeted by the mention (it may be a brand, a product or a feature) must also be identified. Thus, sentiment analysis at aspect-level is deemed more complex than the two above levels. Topic modeling was commonly applied to handle both problems at the same time (Qiu *et al.*, 2011). Mei *et al.* (2007) proposed using Probabilistic Latent Semantic Analysis, while most recent works were based on

Latent Dirichlet Allocation (LDA) (Li *et al.*, 2010; Zhao *et al.*, 2010; Sauper *et al.*, 2011; Mukherjee and Liu, 2012). However, insight analysis (Titov and McDonald, 2008) showed that even though LDA-based approaches may be good at recognizing entities, they are not suitable to infer features of those entities, since all entities of the same domain (e.g. smartphone) would have the same features (e.g. battery, screen, design, etc.).

Discussion

Among the three levels of sentiment analysis, the work presented in this paper falls into the third kind of aspect-based level, since we may want to identify sentiment opinions of certain brands/products/features. To overcome the drawback of the prominent LDA-based methods, we suggest using ontologies. Semantic information of entities and features and can be precisely described in an ontology in a machine-understandable manner. The ontology construction, in this case, should preferably be automatic and adaptable to multiple domains. This motivated us to present a framework which can tackle the aforementioned problems:

Sentiment Ontology for aspect-based sentiment analysis

To give a formal description of aspect-based sentiment analysis, in this section we first present a definition of Sentiment Ontology. Then, we formally define the problem of aspect-based sentiment analysis based on Sentiment Ontology:

Definition 1 (Sentiment Ontology). A Sentiment Ontology S_O is a pair of $\{C,R\}$; where $C=(C^A,C^S)$ represents a set of concepts, which consists of two elements: C^A is a set of aspect concepts, and C^S is a set of sentiment concepts; $R=(R^T, R^N, R^S)$ represents a set of relationships, which consists of three elements: R^N is a set of non-taxonomic relationships, R^T is a set of taxonomic relationships, R^S is a sentiment relationships. Each concept c_i in C represents a set of objects, or instances, of the same kind, denoted as instance-of(c_i). Each relationship $r_i(c_p, c_q)$ in R represents a binary association between concepts c_p and c_q, and the instances of such a relationship, denoted as instance-of(r_i), are pairs of (c_p, c_q) concept objects. Especially, an instance $r_i^s(a,s)$ in R^S implies a relationship between an aspect $a \in A$ and a sentiment term $s \in S$.

Example 1. The Generic Ontology $G_O=\{(C^A,C^S), (R^T, R^N, R^S)\}$ is a Sentiment Ontology where its components are endowed as the following Figure 2.

Generally speaking, G_O consists of one aspect concept of Thing, whose instances can be any real-life concepts. An instance of Thing can be mentioned by a Sentiment Term, which can be either Positive Term or Negative Term. In this example, G_O does not

C^A = {"Thing" }
C^S = {"Sentiment Term", "Negative Term", "Positive Term"}
R^N = {}
R^T = {subconcept-of("Positive Term", "Sentiment Term"), subconcept-of("Negative Term", "Sentiment Term")}
R^S = {mentioned-by("Thing", "Sentiment Term")}
instances-of("Positive Term") = {"like"}
instances-of("Negative Term") = {"hate"}

Figure 2.
The formal representation of Generic Ontology

present any instance of aspect concept, non-taxonomic nor sentiment relation; whereas two terms "like" and "hate" are example instances of sentiment concepts Positive Term and Negative Term, respectively.

To graphically visualize an ontology, we rely on the idea of T-Box and A-Box (Gruber, 1993). Basically, a T-Box captures the relations between concepts and an A-Box describes instances of concepts. Figure 3 presents the T-Box and A-Box of Generic Ontology G_O.

Example 2. Figure 4 gives an example of an Industry Ontology O_S. The T-Box shows that an Industry may have some Brands. Each Brand can produce many Products and each Product has various Features. All of these concepts are subconcepts of Thing in the Generic Ontology, i.e. they can be mentioned by positive or Negative Terms.

The A-Box of this Industry Ontology shows that this ontology describes concepts in the industry of smartphones. There are two brands, namely S-Brand A and S-Brand B, which produce products of Smartphone A and Smartphone B, respectively. A smartphone product has certain features, such as battery and screen. Though sentiment terms are inherited from the concept of Thing, this Industry Ontology also introduces some new domain-specific sentiment terms in a manner such that durable is a Positive Term with regard to battery whereas small size is a Negative Term with respect to screen.

The formal information of this Industry Ontology can be captured as presented in Figure 5:

Definition 2 (aspect-based sentiment analysis). Given a Sentiment Ontology S_O and a textual statement ϑ, an aspect-based sentiment analysis process will recognize all of instances of sentiment relations R^S of S_O in ϑ.

Example 3. Consider the following statement.

$$\text{(S1) Unlike Smartphone A, Smartphone B has a very durable battery.}$$

With Industry Ontology given in Example 2, in (S1) we have two instances of Product concept (i.e. Smartphone A and Smartphone B) and one instance of Positive Term concept (durable). Thus, sentiment relations between those instances are established accordingly. In this paper, we suggest to use CG to capture all of detected relations in a statement. Using a lightweight NLP technique, one can generate the corresponding CG of this

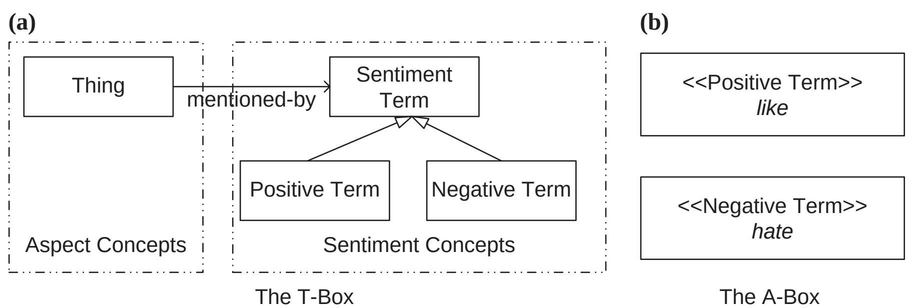

Figure 3.
An example of Generic Ontology

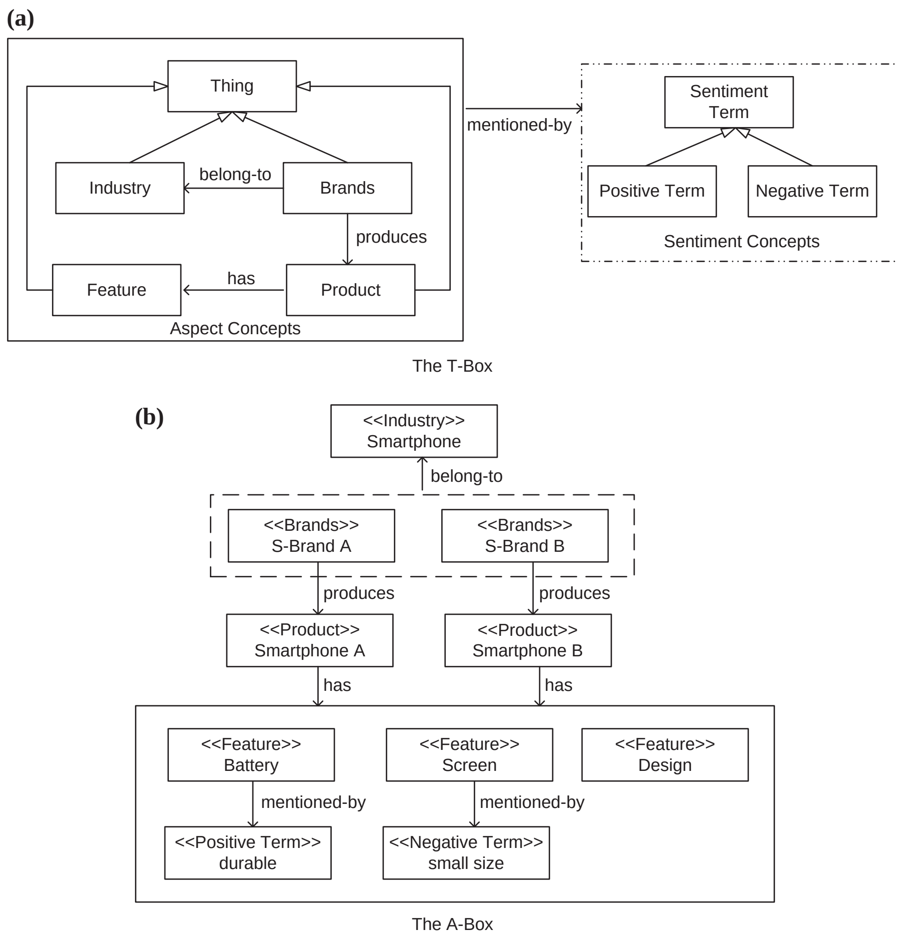

Figure 4.
An example of Industry Ontology

statement as illustrated in Figure 6. From this CG, one can easily infer that Smartphone B is mentioned with positive meaning in this statement, whereas Smartphone A is considered quite negative.

In Quan and Hui (2008), we already presented the technique to construct such a CG from a domain ontology. However, to infer the sentiment opinion, we need to capture more complex linguistic patterns, such as the unlike-phrase given in Example 3. Later in this paper, we will discuss Sentiment Phrasing Rules used to capture such patterns.

CSS: a framework for CSS

Figure 7 presents our CSS framework which is evolved from the common framework previously discussed. To be adapted from a generic semantic search model into the specific sentiment search problem, our CSS framework is enhanced with the following features:

Generic Ontology and Industry Ontologies

In CSS, a common Generic Ontology is constructed which captures domain-independent sentiment terms. Apart from that, a set of domain ontologies are also developed. Since our

//Concepts

C^A = {“Thing”, “Industry”, “Brand”, “Product”, “Feature” }

C^S = {“Sentiment Term”, “Negative Term”, “Positive Term”}

//Relations

R^N = {belong-to(“Brand,Industry”), produce(”Brand”,”Product”), has(“Product”,“Feature”)}

R^T = {subconcept-of(“Industry”,“Thing”), subconcept-of(“Brand”,“Thing”), subconcept-of(“Product”,“ Thing”), subconcept-of(“Feature”,“Thing”), subconcept-of(“Positive Term”,“Sentiment Term”), subconcept-of(“Negative Term”, “Sentiment Term”)}

R^S = {mentioned -by(“Thing”, “Sentiment Term”)}

//Concept instances

instances-of(“Positive Term”) = {“like”,“durable”}

instances-of(“Negative Term”) = {“hate”,“small size”}

instances-of(“Industry”) = {“Smartphone”}

instances-of(“Brands”) = {“S-Brand A”, “S-Brand B”}

instances-of(“Features”) = {“Battery”, “Screen”, “Design”}

//relation instances

instances-of(“produces”) = {(“S-Brand A”,“Smartphone A”), (“S-Brand B”,“Smartphone B”)}

instances-of(“has”) = {(“Smartphone A”,{“Battery”, “Screen”, “Design”}), (“Smartphone A”,{“Battery”, “Screen”, “Design”})}

instances-of(“mentioned-by”) = {(“Battery”,“durable”),(“Design”,“sm all size”)}

Figure 5.
The formal representation of Generic Ontology

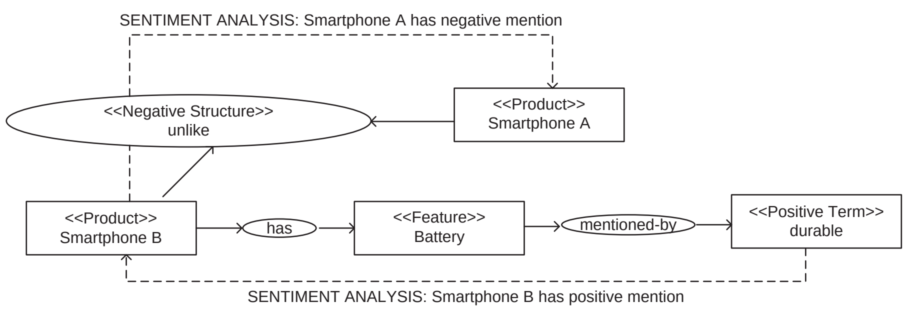

Figure 6.
An example of sentiment analysis on conceptual graph

framework is specifically intended for social monitoring systems which are often used to monitor brands and products in various industries such as Smartphone or Babycares, each domain ontology should capture information in a certain industry. Thus, we regard our domain ontologies as Industry Ontologies. We also develop an incremental strategy to automatically enrich an Industry Ontology from the Generic Ontology. This is introduced as an external module of Incremental Update in CSS.

Sentiment Phrasing Rules

In CSS, we still rely on the lightweight NLP technique of shallow parsing for semantic feature extraction. However, our ontologies define industry concepts and sentiment terms. We develop some specific linguistic rules, known as Sentiment Phrasing Rules, to recognize these concepts and terms in a sentence and generate proper relations between them.

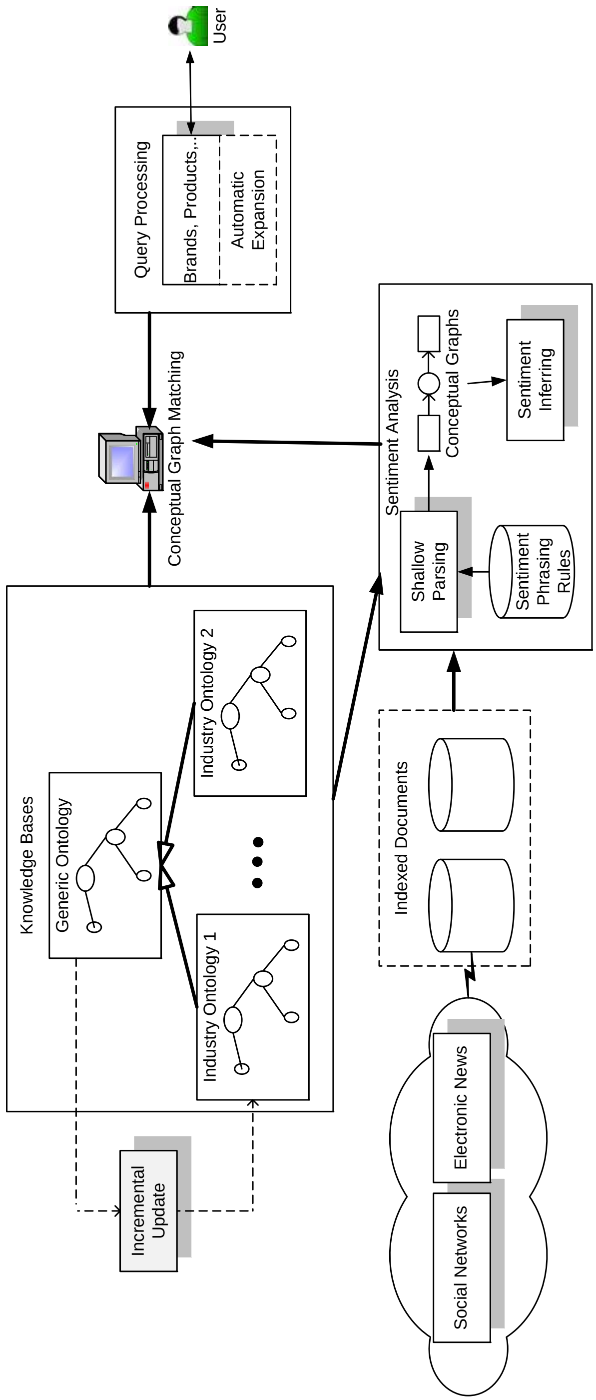

Figure 7.
The Cross-Domain Sentiment Search framework

Query processing
The query processing in CSS is merely keyword-based. The submitted keywords are names of industry, brands or products that users want to observe public sentiment opinions on. First, traditional keyword-based techniques are applied in CSS to retrieve relevant documents. Aspect-based sentiment analysis is then performed to infer the sentiment opinion of the retrieved data.

Sentiment Phrasing Rules

As previously discussed, we develop a set of Sentiment Phrasing Rules capturing linguistic phrases that imply sentiment opinions. The structure of a Sentiment Phrasing Rule is as follows:

```
Sentiment_Phrasing_Rule
#pattern: the pattern of the sentiment phrase captured by these rules.
#aspect: the aspects targeted by sentiment opinion.
#sent_parts: the parts of the phrase expressing the sentiment.
#core_part: the part that expresses the main sentiment trend in phrases.
#neg: flag to indicate whether it is a negative phrase or not.
```

Example 4. Let us consider the following rule:

```
Example_Sentiment_Rule_1
#pattern: (\S+/N\s+)+(\S+/V\s+)+(\S+/A\s*)+
#aspect: N
#sent_parts: [V,A]
#core_part: V
#neg: 0
```

The #pattern of the rule is described by a regular expression (RE), conforming to the RE convention specified at http://regexpal.com/. Roughly speaking, one can read this rule as follows: "This rule applies for the sentence matching the following pattern: There is a noun *N* in the sentence, then a verb *V* after *N*, and then an adverb *A* after *V*."

The #sent_parts specifies that only V and A are necessary to infer the sentiment, whereas, N is the aspect targeted by the sentiment opinion and #core_part specifies that the main sentiment of this phrase can be inferred by V (A will only be taken into account if we are unsure about the sentiment implication of V).

Example 5. The pattern of Example_Sentiment_Rule_1 matches with the following sentence.

The battery of Smartphone A always runs out very fast.

In this sentence, N is battery of Smartphone A, V is run out and A is very fast. According to #sent_parts, we only use V and A to infer the sentiment opinion of this sentence. In this case, V (run out) has negative meaning, meanwhile A (fast) is a positive adverb. Since #sent_parts of this rule states that V will dominate A in terms of sentiment, we conclude that this sentence implies negative meaning over the battery of Smartphone A (the aspect represented by N).

In practice, our Sentiment Phrasing Rules are more complex and cover several cases of sentiment implications. One can obtain our full set of rules consisting of 64 Sentiment Phrasing Rules at www.cse.hcmut.edu.vn/~save/SentimentPhrasing

Rules.pdf. Generally, our rules cover the following issues of sentiment phrases, some of which are referred from the work of Liu (2012):

- Explicit opinion, or regular opinion, e.g., "The picture quality is great."
- Implicit opinion, e.g., "I bought the mattress a week ago, and a valley has formed" (speaker wanted to indicate negative meaning about the mattress).
- Comparative opinion, e.g., "Coke tastes better than Pepsi" (speaker implies positive meaning of Coke and negative meaning of Pepsi).
- Sentiment shifters: the words that can affect the sentiment orientation, including negations (e.g. "no," "not," "never," etc.), intensifiers and diminishers (e.g. "very," "barely," etc.).
- Syntactic dependency: we also refer to the works of Vinodhini and Chandrasekaran (2012) to analyze some special syntax structures that can imply sentiment opinion (e.g. unlike structure as illustrated in Example 3).

Incremental strategy to enrich sentiment terms of Industry Ontology

One of the major problems of sentiment analysis is cross-domain adaptation. For example, twisted plot may imply positive meaning in the domain of Book, but it is virtually meaningless in the domain of Smartphone. Producing domain-specific list of sentiment terms would be a very tedious and highly effort-consuming task, if performed manually. In our approach, we use domain ontology to capture domain knowledge. Our approach is based on our previous work of FOGA framework (Quan *et al.*, 2006) to generate ontology from a data set. In this framework, ontology is to be produced in an offline manner, i.e., the ontology cannot be updated real-time during the retrieval process. However, once ontology is constructed completely, it can be used for retrieval effectively.

Nevertheless, though ontological concepts and instances can be learned effectively by FOGA, the aspect-associated sentiment terms still needed to be built up. In this research, we introduce an incremental strategy to produce such sentiment terms for an Industry Ontology, starting from a generic domain-independent sentiment terms of Generic Ontology.

In order to achieve this goal, we first recall the concept of link strength in information theory (White and Griffith, 1981) as follows:

Definition 3 (Link Strength). Given two term A and B and a document set S_D, the link strength of A and B is calculated using the following formula:

$$Link\_Strenght(AB) = X/(Y - X)$$

where X is the number of documents in S_D in which A and B co-occur and Y is the total number of documents in S_D in which either A or B occurs.

Definition 4 (Link Strength Matrix). Given a set S of N terms $T_{1,}T_2,\ldots, T_N$, the Link Strength Matrix of S is the matrix M of size $N \times N$ where $M(i,j) = Link\_Strength(T_i,T_j)$.

Next, we consider correlation, which is defined as a measure of similarity. The higher the correlation, the more similar the two corresponding terms are. The Pearson's correlation coefficient is commonly used in data analysis to measure the similarity between items in a set (Rodgers and Nicewander, 1988). It is defined as follows:

Definition 5 (Correlation). Given a set of N terms $T_1, T_2, \ldots, T_N$, and their Link Strength Matrix M, the correlation between two terms T_i and T_j is defined as:

$$corr(i,j) = \frac{N(\sum XY) - (\sum X)(\sum Y)}{\sqrt{\left|N\left(\sum X^2 - (\sum X)^2\right)\right| - \left|N\left(\sum Y^2 - (\sum Y)^2\right)\right|}}$$

where X and Y are the ith and the jth rows of M, respectively.

Using the correlation, we can infer the relatedness of generic terms and domain-specific terms. Then, the sentiment terms of an Industry Ontology can be incrementally updated using the algorithm presented below. Furthermore, we use Justeson and Katz's method (Justeson and Katz, 1995) for multi-word extraction to combine sentiment terms into longer sentiment phrases. This conveys more precise meaning when applied in specific domains.

Algorithm Incremental_Update_Domain_Ontology

Input: Domain ontology O_D, generic ontology O_G, a document sets D in the domain of O_D

Process:

repeat
    using O_D and O_G detect sentiment documents D_{sent} (both positive and negative) in D
    detect terms T in D_{sent} which have high correlation with sentiment terms of O_D and O_G
    apply Sentiment Phrasing Rules to infer sentiment orientation of T
    detect *aspect concept* C in O_D which have high correlation with T
    update T as a new instance of sentiment concepts in O_D
until there is not update made for O_D

Output: Updated domain ontologyO_D

Remark 1. The complexity of Algorithm Incremental_Update_Domain_Ontology is bounded by $|D| \times max\_term$, where *max_term* implies the possible maximum number of sentiment terms in a domain, which should be finite. Thus, the algorithm complexity is polynomial, therefore it is decidable and traceable by computer programs. However, due to the large sizes and lengths of possible documents in D, it does not afford to be computed in a real-time manner.

Example 6. Let us consider the three mentions as follows:

- (M1) I do not like Smartphone A.
- (M2) Frankly speaking, I like the eye-catching design of Smartphone B.
- (M3) Even though some say that the design of Smartphone B is a clone, I found it eye-catching and innovative.

Assume that initially we have Generic Ontology O_G and Industry Ontology O_S as described in Examples 1 and 2. Then, O_S will be updated in the following rounds.

Round 1

Intuitively, using O_G and O_S to classify the mentions, we would obtain positive mentions as {M2:(SmartphoneB.design, +)} and negative mentions as {M1:(Smartphone A,-)}

based on the sentiment term of like defined in O_G; SmartphoneB.design and Smartphone A are two instances of aspect concepts detected; symbol + and − denote positive and negative orientation, respectively.

Analyzing M1, we do not detect new sentiment words. Meanwhile, analyzing M2, we detect new sentiment word of eye-catching since this word supposedly has high correlation with like. Eye-catching is then considered as a new positive sentiment word of O_S. This new sentiment word should be associated with the aspect of design since eye-catching and design have a high correlation value.

Round 2

Now as eye-catching is newly recognized as a new sentiment term, we incrementally detect M3 as a positive mention of {M3:(SmartphoneB.design, +)} when performing re-classification. Using the correlation calculation again, we detect new sentiment terms of innovative and clone; where innovative is also considered as a new Positive Term for feature SmartphoneB.design of O_S. Moreover, when applying Sentiment Phrasing Rule (Line 4 of Algorithm Incremental_Update_Domain_Ontology), the system infers that clone may have opinion orientation opposing to that of eye-catching due to the presence of the phrase although. Thus, clone is considered as a Negative Term of O_S. (In fact, in the domain of Smartphone, a product is always criticized if its design is cloned from another product's design).

Round 3

Since there is a no new classification result update, the update of O_S finishes:

Example 7. Let us consider the three mentions as follows:

- (M4) Software S1, which is a clone of software S2 adapted for Vietnamese language, is attracting much attention in the market.
- (M5) The layout of S1 is responsive. I like it.
- (M6) Smartphone A has an eye-catching design but I hate it.

In this example, we suppose that the system already classified mention M4 to the Software Industry and M5 and M6 to the Smartphone Industry. We also assume that the phrases clone and eye-catching have already been incrementally added as negative and Positive Terms, respectively, for the Smartphone Industry, as illustrated in Example 6.

In mention M4, since clone is only considered as a Negative Term for Smartphone Industry, it is considered as a neutral term in this Software Industry mention. Hence, due to the generic Positive Term attract(ing), M4 is regarded as a positive mention of product Software S1, denoted as {M4:(Software S1, +)}. Subsequently, M5 is inferred as a positive mention of Smartphone A due to the Positive Term like. Moreover, new Positive Term responsive is detected. In mention M6, even though the Smartphone-specific Positive Term eye-catching occurs, the Sentiment Phrasing Rule corresponding to but clause decides that the Negative Term hate prevails. Therefore M6 is concluded as a negative mention for Smartphone A.

Case study

In this section, we would like to present a case study of aspect-based sentiment search. In March 2014, Samsung had launched a promotion campaign for their new product Galaxy S5 smartphone on various social media channels (Facebook, hi-tech magazines and forums, etc.) (www.androidcentral.com/samsung-galaxy-s5-launched-australia-pre-orders-start-march-27). In this campaign, the product monitoring team had to monitor user feedbacks for this product, especially negative comments. Additionally, when notified about a positive/negative comment about the product, one would desire to have information about the product feature targeted by this positive/negative comment.

Our CSS framework had been applied to target this problem. We first retrieve all mentions on social media channels which are related to the keyword Samsung Galaxy S5. Then, sentiment analysis can be applied to filter negative mentions about this product. In serious cases, an alarm is sent to the monitoring operators to notify about a potential crisis, as illustrated in Figure 8. Additionally, due to the aspect-level analysis technique applied, the system is capable of recognizing product features being commented upon as positive, negative or neutral sentiment opinions, as shown in Figure 9. The performance of the proposed framework in this case study is presented in the next section.

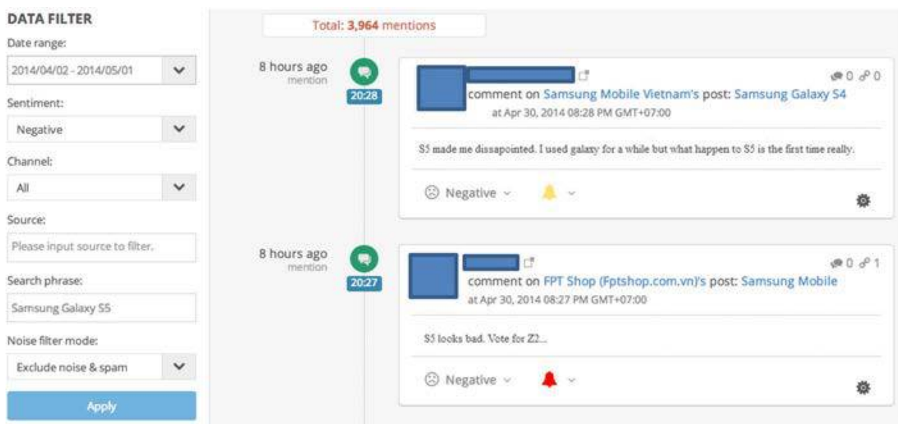

Figure 8.
Monitoring negative mentions of Galaxy S5 product

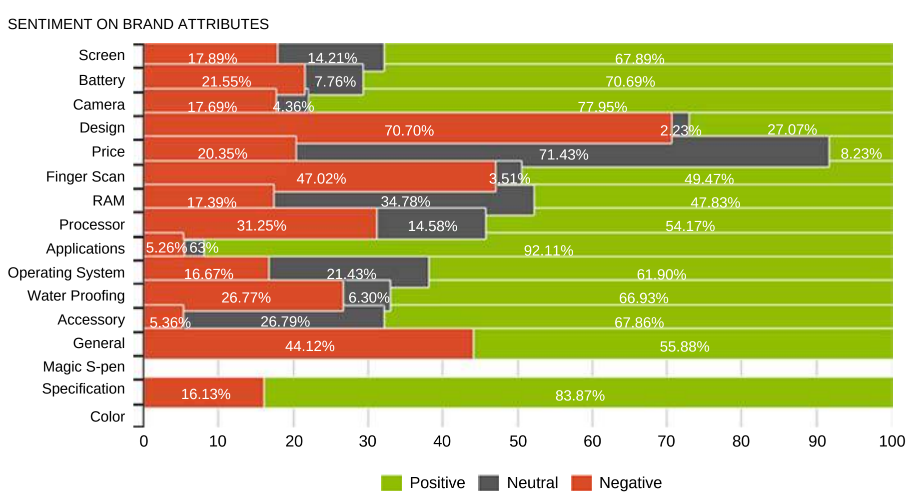

Figure 9.
Feature-based sentiment analysis of Galaxy S5 product

Experimental results

Data sets

To verify the efficiency of our ontological approach, we conducted experiments from real data. The data sets were provided by YouNet Media, a company dedicated to social monitoring systems, in which sentiment search plays an essential role (www.younetmedia.com). Thus, we acquired four different data sets consisting of documents and posts collected from popular electronic newspapers and social networks (hereafter regarded as mentions since they must be mentioning a brand or product when collected). The sentiment implications of those mentions were manually classified by YouNet Media staff.

The data sets were also pre-processed to be standardized and noise-free. The pre-processing consisted of the following tasks:

- *Stemming*: it converts the extracted words to their root forms. For example, "retrieved," "retrieving" and "retrieves" are converted to "retrieve."
- *Stop word removal*: it removes words with weak or no meanings such as "to," "the," "a," etc. Such words are identified through a stop word list.
- *Noise filtering*: it removes spam and advertisement mentions, especially with those from social networks and forums.

To implement these tasks, we used the built-in functions supported by the Solr (http://lucene.apache.org/solr/), which also serves as the core search engine of our system. We also used SVM technique for the noise filtering task.

Alongside with the mention data sets, we produce the corresponding ontologies. As previously discussed, for each industry, we produce the corresponding Industry Ontology capturing major brands and products in the industry. For example, for Smartphone Ontology, the brands included Apple, HTC, etc. and the products included iPhone 5s, iPad Air, HTC One Max, HTC Sensation, etc. Thus, the data sets and ontologies reflected real demands of brand managers who always want to observe the opinion of users regarding their products.

The products are also associated with a set of corresponding features (e.g. features of product Smartphone consists of screen, battery, camera, design, etc. as illustrated in Figure 9). Each Industry Ontology also includes a set of sentiment terms and phrases commonly used for this industry. Besides, all Industry Ontologies also inherited the set of generic sentiment terms stored in Generic Ontology. The set of sentiment terms and phrases in Generic Ontology are developed manually by scanning a dictionary. At the end, we have collected 1,154 Positive Terms and 1,354 Negative Terms. The set of brands, products and features in Industry Ontologies are extracted from various online sources. To develop the sets of specific sentiment terms in Industry Ontology, we first develop small-scale initial sets using manual approaches and then enrich these initial sets as described in Algorithm Incremental_Update_Domain_Ontology.

Table II presents details of our data sets, whose full ontologies can be downloaded from www.cse.hcmut.edu.vn/~save/ontologies.zip

Classification methods

We measured the accuracy of our sentiment classification approach. As compared to traditional framework of generic semantic search, we enhanced our sentiment search by introducing our Sentiment Phrasing Rules and enriching the Industry Ontology.

Therefore, we did compare the performance of various sentiment analysis strategies as follows:

- CSS_{FULL}: we applied our full CSS framework.
- CSS_{GEN}: we only used Generic Ontology in the CSS framework.
- $CSS_{NO\text{-}RULES}$: we did not use Sentiment Phrasing Rules in the CSS framework.
- LDA: we applied the topic model approach, supported by LDA technique.
- SVM: we used SVM for sentiment classification, as this technique was employed by various related works. The latest results of Delta tf.idf metrics (Martineau and Finin, 2009) were also used to obtain the best performance of the SVM technique.

Sentiment classification accuracy

Our sentiment analysis performance is evaluated based on the detection of non-neutral mentions (i.e. negative and positive cases) from the data sets. Obviously, the set of sentiment terms (both positive and negative) plays a crucial role on this task. If no sentiment terms are used, chances are that we cannot detect any non-neutral cases. Nevertheless, when we use the full set of sentiment terms, we can detect possibly maximal numbers of non-neutral cases However, it also increases the chance of false positive detection (i.e. a neutral mention is classified as a positive or negative one).

Hence, in this experiment, we vary the size of sentiment term set from empty to full size. Then, we evaluate the sentiment analysis performance accordingly at each changing point. The results are then represented by the corresponding ROC curves as presented in Figure 10. As observed, all of sentiment analysis techniques used in our experiments achieve relatively good results as the areas covered by their ROC curves are remarkably larger than the value of 0.5 (i.e. the area covered by a random classifier). CSS_{FULL} and LDA generally get better performance the rest of three other methods. LDA is slightly better in terms of true positive detection, but CSS_{FULL} makes less mistake of recognizing false positive cases.

Aspect-based sentiment retrieval efficiency

In this section, we evaluate the retrieval efficiency of our approach while performing aspect-based sentiment retrieval. We have two levels of aspects as follows:

- *Entity-level aspect*: in this level, an aspect is considered as named entities, which are proper names. These proper names could be product names, such as Galaxy S5 or HTC (in domain of Smartphone), or company names such as Techcombank or Sacombank (in the domain of Bank).

Data set	No. of brands	No. of products	Additional terms Positive	Additional terms Negative	No. of mentions Positive	No. of mentions Neutral	No. of mentions Negative
Babycares	100	145	1,142	527	8,013	10,627	4,116
Insurance	49	93	1,975	1,116	5,362	16,227	10,219
Banking	61	67	1,246	875	3,016	22,670	6,725
Smartphones	32	234	1,782	1,469	7,926	10,212	10,411

Table II. Details of data sets used in experiments

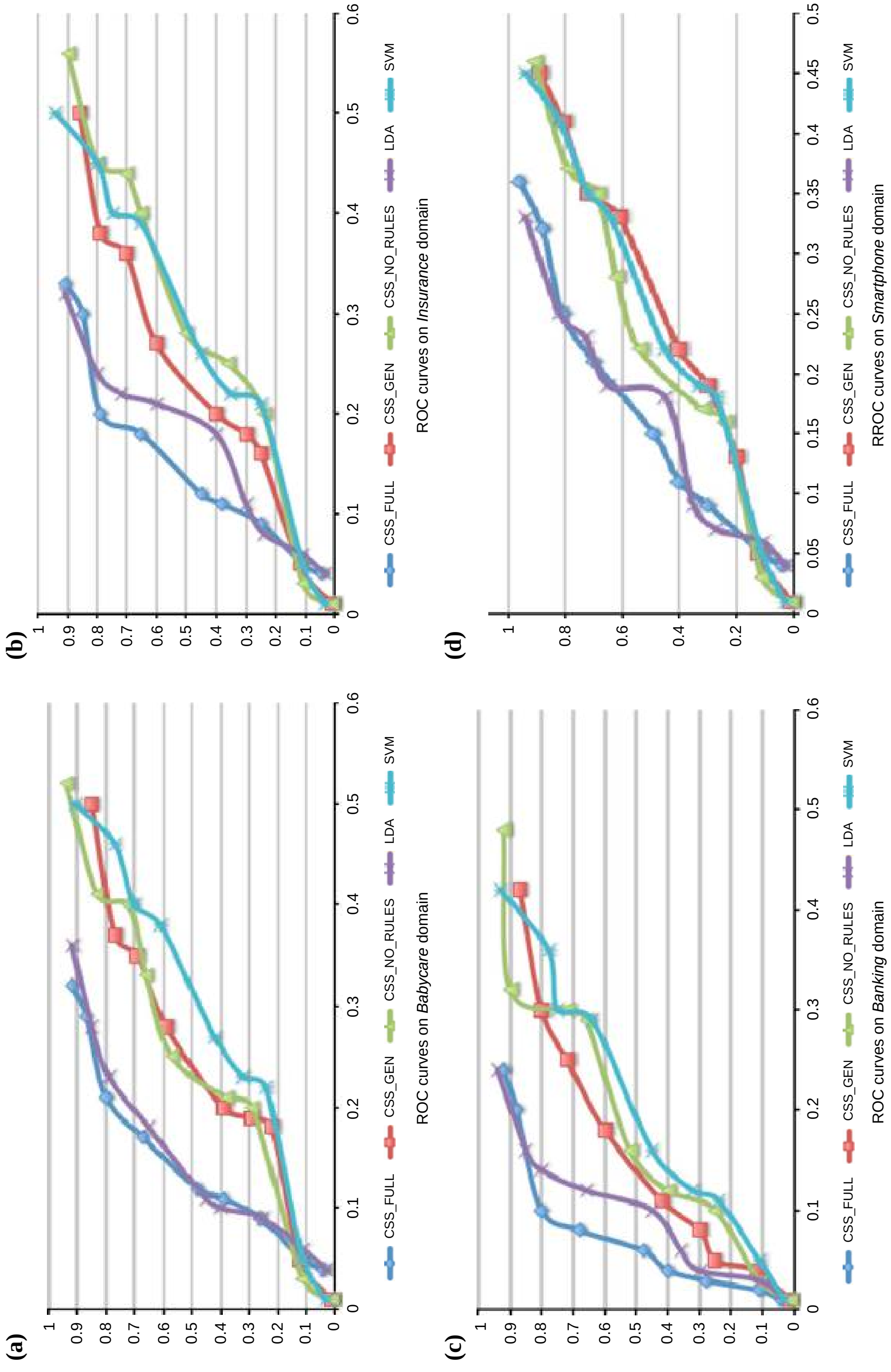

Figure 10.
Accuracy performance of sentiment analysis approaches

- *Feature-level aspect*: in this level, an aspect is a feature of a product or brand, such as pricing or design for smartphone products; or interest rate or customer service for bank companies.

Obviously, feature-level aspects pose more difficulties for processing than entity-level ones. It is because though named entities distinguish themselves from one to another (i.e. two different products/brands will have different names), products or brands of the same domains may have the same set of features (e.g. feature design can be applied for all products in the smartphone domain).

In our aspect-based sentiment retrieval model, each query consists of two parts: an aspect and a desired sentiment orientation. For example, {"S5", "positive"} aim at finding all positive mentions of the product S5, however, {("design of S5"),"negative"} implies finding all mentions which give negative comments about the design of S5.

We use the typical retrieval metrics of recall, precision and F_1 (or F-measure) (Rijsbergen, 1979), to evaluate the efficiency of the retrieval methods. Tables III and IV present the retrieval results when measured at entity-level aspect and feature-level aspect.

The experimental results show that the recall significantly reduces when we use Generic Ontology with CSS_{GEN}. This is because Generic Ontology does not capture the domain-specific terms. Based on probabilistic and machine learning theory, LDA and SVM maintain good recall performance, like what CSS_{FULL} and $CSS_{NO\text{-}RULES}$ achieve when using appropriate Industry Ontologies. However, precision performance of SVM is somehow limited since this method cannot capture linguistic structures. Due to the capability of inferring hidden relationships between words/phrases, LDA enjoys good performance in terms of precision. As a result, CSS_{FULL} and LDA are the two methods that achieve best F_1 results when performing sentiment retrieval at entity-level aspects. Further, it can also be observed that the precisions of all methods are slightly improved when processing on the domain where neutral data is dominant, like Banking.

When handling feature-level aspect, the overall performances of all methods are decreased. LDA, suffering from difficulties when linking features with correct products/brands, also witnesses reducing precision performance. Meanwhile, supported by Sentiment Phrases Rules and products names captured in Industry Ontologies, CSS_{FULL} still gains relatively good precision performance, bringing best F_1 performance achieved by this method.

Conclusion

In this paper we discuss about sentiment search, which can be considered as a special case of semantic search, where the "semantics" is the sentiment or opinion. Aiming specifically at identifying opinions of users when mentioning brands and products in social monitoring systems, we focus on the aspect-based sentiment classification problem and propose a retrieval framework known as CSS. In the proposed framework, domain ontologies are combined with specific linguistic rules, known as Sentiment Phrasing Rules, to efficiently support sentiment search in targeted industries. Additionally, CSS supports incrementally enriching ontologies when applied on new domains. Our experiments on real data sets acquired from actual social channels have achieved promising performance.

	Babycares			Insurance			Banking			Smartphone		
	Recall	Precision	F_1	Recall	Precision	F_1	Recall	Precision	F_1	Recall	Precision	F_1
CSS_{FULL}	92.83	83.03	87.65	91.63	83.44	87.34	92.53	88.28	90.35	94.18	84.13	88.87
CSS_{GEN}	85.71	75.96	80.54	86.06	75.05	80.17	87.64	79.59	83.42	89.52	75.23	81.75
$CSS_{NO\text{-}RULES}$	94.72	74.63	83.48	90.98	72.89	80.93	92.80	76.58	83.91	91.62	74.86	82.39
LDA	92.73	82.24	87.17	90.96	83.34	86.98	93.45	88.28	90.79	94.92	83.74	88.98
SVM	91.00	75.39	82.46	93.69	75.57	83.66	93.13	79.42	85.73	94.07	75.94	84.03

Table III.
Sentiment retrieval performance at entity-level aspects

Table IV.
Sentiment retrieval performance at feature-level aspects

	Babycares			Insurance			Banking			Smartphone		
	Recall	Precision	F_1	Recall	Precision	F_1	Recall	Precision	F_1	Recall	Precision	F_1
CSS_{FULL}	90.48	82.07	86.07	88.79	82.38	84.35	90.33	87.26	88.76	93.12	82.41	87.43
CSS_{GEN}	83.40	69.27	75.68	86.11	74.50	75.97	86.05	73.12	79.05	85.50	71.98	78.15
$CSS_{NO\text{-}RULES}$	92.53	65.38	76.62	88.21	71.59	71.58	90.29	67.92	77.52	88.85	65.53	75.42
LDA	91.06	74.94	82.21	89.82	80.80	79.69	91.69	82.49	86.84	93.31	77.61	84.73
SVM	88.72	66.39	75.94	91.56	73.02	75.01	91.96	72.39	81.00	92.81	67.50	78.15

Notes

1. www.ninesights.com/docs/DOC-8380?utm_source = Email&utm_medium = email&utm_campaign = 69987
2. www.google.com
3. www.wolframalpha.com
4. www.socialmention.com
5. www.jivesoftware.com

References

Abney, S. (1991), "Parsing by chunks", in Berwick, R., Abney, S. and Tenny, C. (Eds), *Principle-Based Parsing*, Kluwer Academic Publishers, Dordrecht, pp. 257-278.

Asher, N., Benamara, F. and Mathieu, Y. (2008), "Distilling opinion in discourse: a preliminary study", in Scott, D. and Uszkoreit, H. (Eds), *Proceedings of the International Conference on Computational Linguistics*, *Coling 2008 Organizing Committee, Manchester*, pp. 7-10.

Auer, S., Dietzold, S. and Riechert, T. (2006), "OntoWiki – a tool for social, semantic collaboration", in Cruz, I.F., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M. and Aroyo, L. (Eds), *The Semantic Web – ISWC 2006, Proceedings of the 5th International Semantic Web Conference, ISWC 2006, Lecture Notes in Computer Science Vol. 4273, Athens, GA, November 5-9, Springer, Heidelberg*, pp. 736-749.

Berkley, C. (2009), "Improving data discovery for metadata repositories through semantic search", in Barolli, L., Xhafa, F. and Hsu, H.H. (Eds), *Proceedings of the 2009 International Conference on Complex, Intelligent and Software Intensive Systems, CISIS 2009, Fukuoka, March 16-19, IEEE Computer Society, Piscataway*, pp. 1152-1159.

Bhagdev, R., Chapman, S., Ciravegna, F., Lanfranchi, V. and Petrelli, D. (2008), "Hybrid search: effectively combining keywords and semantic searches", in Bechhofer, S., Hauswirth, M., Hoffmann, J. and Koubarakis, M. (Eds), *The Semantic Web: Research and Applications, Proceedings of the 5th European Semantic Web Conference, ESWC 2008, Lecture Notes in Computer Science, Vol. 5021, Tenerife, Canary Islands, June 1-5, Springer*, pp. 554-568.

Bottou, L. (2011), "From machine learning to machine reasoning", *Computing Research Repository*, Vol. abs/1102.1808, available at: http://arxiv.org/pdf/1102.1808v3.pdf

Bruza, P., Koopman, B., Lawley, M. and Sitbon, L. (2012), "Towards semantic search and inference in electronic medical records: an approach using concept-based information retrieval", *Australas Med J*, Vol. 5 No. 9, pp. 482-488.

Cimiano, P., Rudolph, S., Studer, R. and Tran, T. (2007), "Ontology-based interpretation of keywords for semantic search", in Aberer, K., Choi, K.S., Noy, N.F., Allemang, D., Lee, K.I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G. and Cudré-Mauroux, P. (Eds), *The Semantic Web, Proceedings of the 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Lecture Notes in Computer Science, Vol. 4825, Busan, November 11-15, Springer, Heidelberg*, pp. 523-536.

Cohen, S., Mamou, J., Kanza, Y. and Sagiv, Y. (2003), "XSEarch: a semantic search engine for XML", in Freytag, J.C., Lockemann, P.C., Abiteboul, S., Carey, M.J., Selinger, P.G. and Heuer, A. (Eds), *Proceedings of the 29th International Conference on Very large Data Bases, ACM, New York, NY*, pp. 45-56.

Dave, K., Lawrence, S. and Pennock, M. (2003), "Mining the peanut gallery: opinion extraction and semantic classification of product reviews", in Hencsey, G., White, B., Chen, Y.F.R., Kovács, L. and Lawrence, S. (Eds), *Proceedings of the 12th International World Wide Web Conference, WWW 2003, Budapest, May 20-24, ACM, New York, NY*, pp. 519-528.

Duboue, P., Czuba, K., Chu-Carroll, J., Ferrucci, D. and Prager, J. (2006), "Semantic search via XML fragments: a high-precision approach to IR", in Efthimiadis, E.N., Dumais, S.T., Hawking, D. and Järvelin, K. (Eds), *SIGIR 2006: Proceedings of the 29th Annual International ACM*

SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, DC, August 6-11, ACM, New York, NY, pp. 445-452.

Dumais, S.T. (2004), "Latent semantic analysis", *Annual Review of Information Science and Technology*, Vol. 38 No. 1, pp. 188-230.

Dietze, H. and Schroeder, M. (2009), "GoWeb: a semantic search engine for the life science web", *BMC Bioinformatics*, Vol. 10 No. S10, p. S7.

Embley, D., Liddle, S., Lonsdale, D. and Tijerino, Y. (2011), "Multilingual ontologies for cross-language information extraction and semantic search", in Jeusfeld, M.A., Delcambre, L.M.L. and Ling, T.W. (Eds), *Conceptual Modeling – ER 2011, Proceedings of the 30th International Conference, ER 2011, Lecture Notes in Computer Science, Vol. 6998, Brussels, October 31-November 3, Springer, Heidelberg*, pp. 147-160.

Fernandez, M., Lopez, V., Sabou, M., Uren, V.S., Vallet, D., Motta, E. and Castells, P. (2008), "Semantic search meets the web", in Sheu, P., Croll, P., Mühlhäuser, M., Ishizuka, M., Hyvönen, E., Palmer, M. and Heather, Y.H. (Eds), *Proceedings of the 2th IEEE International Conference on Semantic Computing (ICSC 2008), Santa Clara, CA, August 4-7, IEEE Computer Society, Piscataway*, pp. 253-260.

Ferré, S. and Hermann, A. (2011), "Semantic search: reconciling expressive querying and exploratory search", in Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N.F. and Blomqvist E. (Eds), *The Semantic Web – ISWC 2011 – Proceedings of the 10th International Semantic Web Conference, Lecture Notes in Computer Science, Vol. 7031, Part I, Bonn, Germany, October 23-27, Springer, Heidelberg*, pp. 177-192.

Gindl, S. and Liegl, J. (2008), "Evaluation of different sentiment detection methods for polarity classification on web-based reviews", in Ghallab, M., Spyropoulos, C.D., Fakotakis, N. and Avouris, N. (Eds), *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI-2008), ECAI Workshop on Computational Aspects of Affectual and Emotional Interaction, 18th European Conference on Artificial Intelligence, Patras, IOS Press, Amsterdam*, pp. 35-43.

Glorot, X., Bordes, A. and Bengio, Y. (2011), "Domain adaptation for large-scale sentiment classification: a deep learning approach", in Getoor, L. and Scheffer, T. (Eds), *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, DC, June 28-July 2, Omnipress, Madison*, pp. 513-520.

Go, A., Bhayani, R. and Huang, L. (2009), "Twitter sentiment classification using distant supervision", CS224N Project Report, Stanford, pp. 1-12.

Gruber, R. (1993), "A translation approach to portable ontology specifications", *Knowledge Acquisition*, Vol. 5 No. 2, pp. 199-220.

Guha, R.V., McCool, R. and Miller, E. (2003), "Semantic search", in Hencsey, G., White, B., Chen, Y.F.R., Kovács, L. and Lawrence, S. (Eds), *Proceedings of the 12th International World Wide Web Conference, WWW 2003, Budapest, Hungary, May 20-24, ACM, New York, NY*, pp. 700-709.

Haase, P., Herzig, D.M., Musen, M.A. and Tran, T. (2009), "Semantic Wiki search", in Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M. and Simperl, E.P.B. (Eds), *The Semantic Web: Research and Applications, Proceedings of the 6th European Semantic Web Conference, ESWC 2009, Lecture Notes in Computer Science, Vol. 5554, Heraklion, Crete, May 31-June 4, Springer, Heidelberg*, pp. 445-460.

Hazewinkel, M. (Ed.) (2001), "Bayesian approach to statistical problems", *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4a.

Hinton, G.E. (1990), "Mapping part-whole hierarchies into connectionist networks", *Artificial Intelligence*, Vol. 46 Nos 1/2, pp. 47-75.

Hu, M. and Liu, B. (2004), "Mining and summarizing customer reviews", in Kim, W., Kohavi, R., Gehrke, J. and DuMouchel, W. (Eds), *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, DC, August 22-25, ACM, New York, NY*, pp. 168-177.

Hyvönen, E., Saarela, S. and Viljanen, K. (2004), "Application of ontology techniques to view-based semantic search and browsing", in Bussler, C., Davies, J., Fensel, D. and Studer, R. (Eds), *The Semantic Web: Research and Applications, Proceedings of 1st European Semantic Web Symposium, ESWS 2004,Lecture Notes in Computer Science, Vol. 3053, Heraklion, May 10-12, Springer, Heidelberg*, pp. 92-106.

Justeson, J.S. and Katz, S.M. (1995), "Technical terminology: some linguistic properties and an algorithm for identification in text", *Natural Language Engineering*, Vol. 1 No. 1, pp. 9-27.

Kandogan, E., Krishnamurthy, R., Raghavan, S., Vaithyanathan, S. and Zhu, H. (2006), "Avatar semantic search: a database approach to information retrieval", in Chaudhuri, S., Hristidis, V. and Polyzotis, N. (Eds), *Proceedings of the ACM SIGMOD International Conference on Management of Data, Chicago, IL, June 27-29, ACM, New York, NY*, pp. 790-792.

Kim, J., Li, J.J. and Lee, J.H. (2009), "Discovering the discriminative views: measuring term weights for sentiment analysis", in Su, K.Y., Su, J. and Wiebe, J. (Eds), *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, August 2-7, The Association for Computer Linguistics, Stroudsburg*, pp. 253-261.

Lei, Y., Motta, E. and Uren, V.S. (2006), "SemSearch: a search engine for the semantic web", in Staab, S. and Svátek, V. (Eds), *Managing Knowledge in a World of Networks, Proceedings of the 15th International Conference, EKAW 2006, Lecture Notes in Computer Science, Vol. 4248, Podebrady, October 2-6, Springer, Heidelberg*, pp. 238-245.

Li, F., Huang, M. and Zhu, X. (2010), "Sentiment analysis with global topics and local dependency", in Fox, M. and Poole, D. (Eds), *Proceedings of the 24th AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, GA, July 11-15, AAAI Press, Palo Alto, CA*, pp. 1371-1376.

Li, J., Yu, Y., Zhong, J. and Zhu, H. (2002), "Conceptual graph matching for semantic search", in Priss, U., Corbett, D. and Angelova, G. (Eds), *Conceptual Structures: Integration and Interfaces, Proceedings of the 10th International Conference on Conceptual Structures, ICCS 2002, Lecture Notes in Computer Science, Vol. 2393, Borovets, July 15-19, Springer, Heidelberg*, pp. 92-106.

Lin, D. (1998), "Automatic retrieval and clustering of similar words", in Boitet, C. and Whitelock, P. (Eds), *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL '98, Université de Montréal, Montréal, Quebec, August 10-14, Morgan Kaufmann Publishers/ The Association for Computer Linguistics, Stroudsburg*, pp. 768-774.

Liu, B. (2012), *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, CA.

Liu, Q., Tran, T., Wang, H., Yu, Y. and Zhang, K. (2008), "Q2Semantic: a lightweight keyword interface to semantic search", in Bechhofer, S., Hauswirth, M., Hoffmann, J. and Koubarakis, M. (Eds), *The Semantic Web: Research and Applications, Proceedings of the 5th European Semantic Web Conference, ESWC 2008, Lecture Notes in Computer Science, Vol. 5021, Tenerife, Canary Islands, June 1-5, Springer*, pp. 584-598.

McDonald, R.T., Hannan, K., Neylon, T., Wells, M. and Reynar, J.C. (2007), "Structured models for fine-to-coarse sentiment analysis", in Carroll, J.A., van den Bosch, A. and Zaenen, A. (Eds), *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, June 23-30, The Association for Computational Linguistics, Stroudsburg*, pp. 432-439.

Martineau, J. and Finin, T. (2009), "Delta TFIDF: an improved feature space for sentiment analysis", in Adar E., Hurst, M., Finin, T., Glance, N.S., Nicolov, N. and Tseng, B.L. (Eds), *Proceedings of the 3rd International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, CA, May 17-20, The AAAI Press*.

Mei, Q., Ling, X., Wondra, M., Su, H. and Zhai, C. (2007), "Topic sentiment mixture: modeling facets and opinions in weblogs", in Williamson, C.L., Zurko, M.E., Patel-Schneider, P.F. and

Shenoy, P.J. (Eds), *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, May 8-12, ACM, New York, NY*, pp. 171-180.

Melville, P., Gryc, W. and Lawrence, R.D. (2009), "Sentiment analysis of blogs by combining lexical knowledge with text classification", in Elder J.F. IV, Fogelman-Soulié, F., Flach, P.A. and Zaki, M.J. (Eds), *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, June 28-July 1, ACM, New York, NY*, pp. 1275-1284.

Miller, G.A., Beckwith, G.A., Fellbaum, C.D., Gross, D. and Miller, K. (1990), "WordNet: an online lexical database", *Int. J. Lexicograph.*, Vol. 3 No. 4, pp. 235-244.

Mukherjee, A. and Liu, B. (2012), "Aspect extraction through semi-supervised modeling", *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers 1, The Association for Computer Linguistics*, pp. 339-348.

Narayanan, R., Liu, B. and Choudhary, A. (2009), "Sentiment analysis of conditional sentences", *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, The Association for Computer Linguistics, Singapore, August 6-7*, pp. 180-189.

Nasukawa, T. and Yi, J. (2003), "Sentiment analysis: capturing favorability using natural language processing", in Gennari, J.H., Porter, B.W. and Gil, Y. (Eds), *Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP 2003), Sanibel Island, FL, October 23-25, ACM, New York, NY*, pp. 70-77.

Padmaja, S. and Fatima, S. (2013), "Opinion mining and sentiment analysis – an assessment of peoples' belief: a survey", *International Journal of Ad hoc, Sensor & Ubiquitous Computing*, Vol. 4 No. 1, pp. 21-33.

Paltoglou, G. and Thelwall, M. (2010), "A study of information retrieval weighting schemes for sentiment analysis", in Hajic, J., Carberry, S. and Clark, S. (Eds), *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, July 11-16, The Association for Computer Linguistics, Stroudsburg*, pp. 1386-1395.

Paolucci, M., Sycara, K. and Srinivasan, N. (2005), "An efficient algorithm for OWL-S based semantic search in UDDI", in Cardoso, J. and Sheth, A.P. (Eds), *Semantic Web Services and Web Process Composition, First International Workshop, SWSWPC 2004, San Diego, CA, Revised Selected Papers, Lecture Notes in Computer Science Vol. 3387, July 6, Springer, Heidelberg*, pp. 96-110.

Prabowo, R. and Thelwall, M. (2009), "Sentiment analysis: a combined approach", *Journal of Informetrics*, Vol. 3 No. 2, pp. 143-157.

Qiu, G., Liu, B., Bu, J. and Chen, C. (2011), "Opinion word expansion and target extraction through double propagation", *Computational Linguistics*, Vol. 37 No. 1, pp. 9-27.

Qu, Y., Qi, G., Wu, K., Wu, G. and Yang, M. (2010), "Falconer: once SIOC meets semantic search engine", in Rappa, M., Jones, P., Freire, J. and Chakrabarti, S. (Eds), *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, April 26-30, ACM, New York, NY*, pp. 1317-1320.

Quan, T.T. and Hui, S.C. (2008), "Ontology-based natural query retrieval using conceptual graphs", in Ho, T.B. and Zhou, Z.H. (Eds), *PRICAI 2008: Trends in Artificial Intelligence, Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence, Lecture Notes in Computer Science, Vol. 5351, Hanoi, December 15-19, Springer, Heidelberg*, pp. 309-320.

Quan, T.T., Hui, S.C., Fong, A.C.M. and Cao, T.H. (2006), "Automatic fuzzy ontology generation for Semantic web", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18 No. 6, pp. 842-884, 856.

Rodgers, J.L. and Nicewander, W.A. (1988), "Thirteen ways to look at the correlation coefficient", *The American Statistician*, Vol. 42 No. 1, pp. 59-66.

Rijsbergen, C.J.V. (1979), *Information Retrieval*, 2nd ed., Butterworths, London.

Salton, G. and Buckley, C. (1988), "Term-weighting approaches in automatic text retrieval", *Information Processing and Management*, Vol. 24 No. 5, pp. 513-523.

Sauper, C., Haghighi, A. and Barzilay, R. (2011), "Content models with attitude", in Lin, D., Matsumoto, Y. and Mihalcea, R. (Eds), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, June 19-24, The Association for Computer Linguistics, Stroudsburg*, pp. 350-358.

Schreiber, G., Amin, A., Aroyo, L., Assem, M., Boer, V., Hardman, L., Hildebrand, M., Omelayenko, B., Osenbruggen, J., Tordai, A., Wielemaker, J. and Wielinga, B. (2008), "Semantic annotation and search of cultural-heritage collections: the multimedian E-culture demonstrator", *Journal of Web Semantics*, Vol. 6 No, 4, pp. 243-249.

Shimada, K. and Endo, T. (2008), "Seeing several stars: a rating inference task for a document containing several evaluation criteria", in Washio, T., Suzuki, E., Ting, K.M. and Inokuchi, A. (Eds), *Advances in Knowledge Discovery and Data Mining, Proceedings of the 12th Pacific-Asia Conference, PAKDD 2008, Lecture Notes in Computer Science, Vol. 5012, Osaka, May 20-23, Springer, Heidelberg*, pp. 1006-1014.

Smith, K.A., Huber, D.E. and Vul, E. (2013), "Multiply-constrained semantic search in the Remote Associates Test", *International Journal of Cognitive Science*, Vol. 128 No. 1, pp. 64-75.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A. and Potts, C. (2013), "Recursive deep models for semantic compositionality over a sentiment treebank", *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pp. 1631-1642.

Somasundaran, S., Ruppenhofer, J. and Wiebe, J. (2008), "Discourse level opinion relations: an annotation study", in Hockey, B.A. and Schlangen, D. (Eds), *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, The Association for Computer Linguistics, Stroudsburg*, pp. 129-137.

Sowa, J.F. (1976), "Conceptual graphs for a data base interface", *IBM Journal of Research and Development*, Vol. 20 No. 4, pp. 336-357.

Täckström, O. and McDonald, R. (2011), "Discovering fine-grained sentiment with latent variable structured prediction models", in Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H. and Murdock, V. (Eds), *Advances in Information Retrieval – Proceedings of the 33rd European Conference on IR Research, ECIR 2011, Lecture Notes in Computer Science, Vol. 6611, Dublin, April 18-21, Springer, Heidelberg*, pp. 368-374.

Tan, S. and Zhang, J. (2008), "An empirical study of sentiment analysis for Chinese documents", *Expert Systems with Applications*, Vol. 34 No. 4, pp. 2622-2629.

Titov, I. and McDonald, R. (2008), "A joint model of text and aspect ratings for sentiment summarization", in McKeown, K., Moore, J.D., Teufel, S., Allan, J. and Furui, S. (Eds), *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, Columbus, OH, June 15-20, The Association for Computer Linguistics, Stroudsburg*, pp. 308-316.

Vinodhini, G. and Chandrasekaran, R. (2012), "Sentiment analysis and opinion mining: a survey", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2 No. 6, pp. 282-292.

Wang, C., Wang, H., Xiong, M., Yu, Y. and Zhou, Q. (2007), "SPARK: adapting keyword query to semantic search", in Aberer, K., Choi, K.S., Noy, N.F., Allemang, D., Lee, K.I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G. and Cudré-Mauroux, P. (Eds), *The Semantic Web, Proceedings of the 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Lecture Notes in Computer Science, Vol. 4825, Busan, November 11-15, Springer, Heidelberg*, pp. 694-707.

White, H.D. and Griffith, B.C. (1981), "Author co-citation: a literature measure of intellectual structure", *Journal of the American Society for Information Studies*, Vol. 32 No. 3, pp. 163-171.

Xia, R., Zong, C. and Li, S. (2011), "Ensemble of feature sets and classification algorithms for sentiment classification", *Information Sciences*, Vol. 181 No. 6, pp. 1138-1152.

Xu, K., Liao, S.S., Li, J. and Song, Y. (2011), "Mining comparative opinions from customer reviews for competitive intelligence", *Decision Support Systems*, Vol. 50 No. 4, pp. 743-754.

Ye, Q., Zhang, Z. and Law, R. (2009), "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches", *Expert Systems with Applications*, Vol. 36 No. 3, pp. 6527-6535.

Yu, H. and Hatzivassiloglou, V. (2003), "Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences", in Collins, M. and Steedman, M. (Eds), *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, The Association for Computer Linguistics, Stroudsburg*, pp. 129-136.

Zhang, Z., Ye, Q., Zhang, Z. and Li, Y. (2011), "Sentiment classification of internet restaurant reviews written in Cantonese", *Expert Systems with Applications*, Vol. 38 No. 6, pp. 7674-7682.

Zhao, W.X., Jiang, J., Yan, H. and Li, X. (2010), "Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid", in Li, H. and Màrquez, L. (Eds), *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, MIT Stata Center, MA, October 9-11, The Association for Computer Linguistics, Stroudsburg*, pp. 56-65.

Zhou, L., Li, B., Gao, W., Wei, Z. and Wong, K.F. (2011), "Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities", in Merlo, P., Barzilay, R. and Johnson, M. (Eds), *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, John McIntyre Conference Centre, Edinburgh, July 27-31, The Association for Computer Linguistics, Stroudsburg*, pp. 162-171.

Zirn, C., Niepert, M., Stuckenschmidt, H. and Strube, M. (2011), "Fine-grained sentiment analysis with structural features", in Wong, K.F., Wang, H. and Yarowsky, D. (Eds), *Proceedings of the 5th International Joint Conference on Natural Language Processing, IJCNLP 2011, Chiang Mai, November 8-13, The Association for Computer Linguistics, Stroudsburg*, pp. 336-344.

Further reading

Liu, B., Hu, M. and Cheng, J. (2005), "Opinion observer: analyzing and comparing opinions on the web", in Ellis, A. and Hagino, T. (Eds), *Proceedings of the 14th International Conference on World Wide Web, WWW 2005, Chiba, May 10-14, ACM, New York, NY*, pp. 342-351.

Wang, H., Lu, Y. and Zhai, C. (2010), "Latent aspect rating analysis on review text data: a rating regression approach", in Rao, B., Krishnapuram, B., Tomkins, A. and Yang, Q. (Eds), *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, July 25-28, ACM, New York, NY*, pp. 783-792.

About the authors

Tung Thanh Nguyen is a Research Student at the Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Vietnam. His research intestests include sentiment analysis, information retrieval and natural language processing.

Tho Thanh Quan is an Associate Professor in the Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Vietnam. He received his BEng Degree in Information Technology from HCMUT in 1998 and received PhD Degree in 2006 from the Nanyang Technological University, Singapore. His current research interests include formal methods, program analysis/verification, the Semantic Web, machine learning/data mining and intelligent systems. Associate Professor Tho Thanh Quan is the corresponding author and can be contacted at: qttho@cse.hcmut.edu.vn

Dr Tuoi Thi Phan is a Full Professor at the Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Vietnam. She is leading a research group on natural language processing at the faculty. Her research interests include machine translation, question answering systems and sentiment analysis.

This article has been cited by:

1. AgostinoDeborah Deborah Agostino Deborah Agostino is based at the Politecnico di Milano, Milano, Italy. She is Assistant Professor at Politecnico di Milano, Department of Management, Economics and Industrial Engineering. Her research interests cover mainly two areas: performance measurement systems in the social media era and performance management in public service networks. Her publications can be found in Public Management Review, European Management Journal and Transport Policy. SidorovaYulia Yulia Sidorova Yulia Sidorova is PhD candidate at the Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Milano, Italy. She is a PhD student at Politecnico di Milano, Department of Management, Economics and Industrial Engineering. Her research concentrates on social media and social media information implementation in management control. The topic is investigated through the action research and multiple case studies. The results of the research have been presented at international conferences. Politecnico di Milano, Milano, Italy Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Milano, Italy . 2016. A performance measurement system to quantify the contribution of social media: new requirements for metrics and methods. *Measuring Business Excellence* **20**:2, 38-51. [Abstract] [Full Text] [PDF]
2. Ya-Ching Lee Institute of Marketing Communication, Kaohsiung, Taiwan . 2016. Determinants of effective SoLoMo advertising from the perspective of social capital. *Aslib Journal of Information Management* **68**:3, 326-346. [Abstract] [Full Text] [PDF]