Ontology-Based Sentiment Analysis Model of Customer Reviews for Electronic Products

K. M. Sam and C. R. Chatwin

Abstract—**This paper reports on a generalizable system model design that analyzes the unstructured customer reviews inside the posts about electronic products on social networking websites. For the purposes of this study, posts on social networking websites have been mined and the keywords are extracted from such posts. The extracted keywords and the ontologies of electronic products and emotions form the base for the sentiment analysis model which is used to understand online consumer behavior in the market. In order to enhance system accuracy, negating and enhancing terms are considered in the proposed model. Sentiment analysis is demonstrated to be extremely important to system accuracy.**

Index Terms—**Semantic content retrieval, sentiment analysis, non-rule based unstructured data, query analysis, HowNet, ontology.**

I. Introduction

Text mining emerged as a means to derive knowledge from unstructured data, especially data available on the World Wide Web. Mine, Lu and Amamiya [1] suggest a text mining system that obtains the relationship between the topics at international conferences. This experiment promises that the method works not only for obtaining the relationship between topics of conferences, but also for discovering the relationship between information entities that users are interested in. The issues regarding text mining have been discussed by Castellano and Mastronardi [2] who have developed a web text mining flexible architecture, which can discover knowledge in a distributed and heterogeneous multi-organization environment. Sennellart and Blondel [3] offer ways for discovery of similar words from the WWW corpus. However both these works were based on text that follows a set of English grammar standard syntax. The mined information is therefore based on predetermined relationships using specific rules and ontologies. However, with the advent of social networking websites, a lot of information is in a non-rule based textual format. The usefulness of this information to determine social behavior is demonstrated by Java, Song, Finin and Tseng [4]. In view of this development, it becomes essential to extract behavioral patterns from relationships established using these data sources rather than predefined associations in ontologies. This paper proposes a mechanism to analyze social behavior through extraction of relationships from data available on online social forums.

Manuscript received March 10, 2013; revised July 30, 2013.

K. M. Sam is with the University of Macau, Macao (e-mail: tonysam@umac.mo).

C. R. Chatwin is with University of Sussex, United Kingdom (e-mail: c.r.chatwin@sussex.ac.uk).

A. Ontology

Ontology is defined as a set of representational primitives used to model a domain of knowledge [5]. A JAVA based ontology editor and knowledge based framework called Protégé [6] was used to develop the ontology, this provides a plug-and-play environment, which is a flexible tool for rapid prototyping and application development.

There is significant ontology-related research based on information retrieval [4], [7], [8], which focuses on two areas: one is to excavate the unknown information using the intrinsic relationship between ontology concepts [9, 10], the other is to use an ontology to classify the searched documents in accordance with the user's personal preferences in order to enhance the efficiency of query [11]. This research aims to offer a new ontology-based solution for semantic content retrieval based on users' queries.

The paper describes an ontology-based sentiment analysis model in detail. Section II discusses the ontology development used for the system. Section III and its subsections discuss the implementation of the system along with a brief description of each module. In Section IV, the experimental results are discussed. Finally, Section V consists of a conclusion about the proposed model.

II. Ontology Development

By using the ontology modeling environment Protégé [6], an ontology primarily based on electronic products has been created. Additionally, the ontology based on emotions and language used on social networking websites is also developed and merged with the electronic products ontology to analyze consumer behavior on social networking websites.

A. Emotions Ontology

The emotions ontology describes the emotions of consumers codifying their feelings related to different products. The ontology starts with the class sentiment, which has the subclasses *happiness* and *sadness*. In order to develop the emotions ontology, a two-step approach is used: i) an online survey has been performed and a sample of 483 customer reviews on electronic products were analyzed to find out the common emotional keywords for electronic products. The keywords under the subclass *happiness* are; *cool*, *great*, *enjoy*, *fantastic*, *excited*, *marvelous*, *perfect*, *love*, *good*, *excellent*, etc. The keywords associated with the subclass *sadness* are: bored, tired, frustrated, dislike, sad, disappointed, bad, worst, etc. ii) the happiness keywords are organized as different levels of subclasses under the happiness class based on the intensity level of happiness. Similarly, the sadness keywords are organized as different

DOI: 10.7763/IJEEEE.2013.V3.282

levels of subclasses under the sadness class based on the intensity level of sadness. The higher the subclass level, the higher the intensity level. In this research, the emotional keywords are mapped to HowNet, an online commonsense knowledge base unveiling inter-conceptual relationships and inter-attribute relations of concepts, in order to determine the similarity among the emotional keywords.

Semantic Similarity

HowNet is different from other semantic dictionaries and sememes are used in HowNet to describe every concept of each term. So the calculation of the semantic similarity has two hierarchies: the semantic similarity calculation of the words switches to calculations of the meanings, and the calculation of the meanings switches to calculations of the sememes.

Suppose there are two words, W_1 and W_2. If W_1 has n meanings $\{S_{11}, S_{12}, \dots, S_{1n}\}$, W_2 has m meanings $\{S_{21}, S_{22}, \dots, S_{2m}\}$, the similarity of W_1 and W_2 can be defined as the maximum similarity of each meaning in (1):

$$Sim(W_1, W_2) = \max_{i=1..n, j=1..m} Sim(S_{1i}, S_{2j}) \qquad (1)$$

Thus, the similarity of two words can be distributed to the similarity of two meanings.

Since most meanings can be represented by sememes, we make the sememes similarity the basis of the meanings similarity. In the dendriform hierarchical system of sememes organized by the Hypernym-Hyponym relation, the most important relation in the sememe relations, the method of calculating the semantic distance is used to get the similarity. The semantic distance $Sim(P_1, P_2) \in [0, 1]$ can be evaluated as (2):

$$Sim(P_1, P_2) = \frac{\alpha}{d+\alpha} \qquad (2)$$

where P_1 *and* P_2 are two sememes, d is the path length in the dendriform-sememes-hierarchy system, α is an adjustment parameter.

It is the basis for calculating the semantic similarity to make up the dendriform hierarchical system of sememes. The semantic similarity based method is a suitable method for text representation [12]. In 2007, HowNet released Chinese/English vocabulary for sentiment analysis. The sentiment information has just been reinforced in HowNet [13] and research about sentiment analysis has been done by using HowNet [14]. Based on the results of the concept similarity calculator shown in Fig. 1, the relevant keywords such as great and marvelous, which shows a similarity of 0.95, are stored in the same subclass level of their corresponding emotional class while the keywords with lower similarity are stored in different subclass levels. Note that the emotional keywords "great" and "marvelous" are organized as subclasses under the same level of the class happiness. Fig. 2 illustrates happiness and sadness classes with part of their sub-classes (including a subclass level number as subscript). The higher the subclass level, the lower the subclass level number starting from 1.

B. Electronic Products Ontology

This ontology aims to provide a controlled vocabulary to semantically describe concepts about electronic products and it also demonstrates the relationships between the various concepts [5]. The ontology starts with the element type of product: computer, computer-related products and household products. Then, their respective sub-categories are built as their sub-classes. The attributes of the product class are summed up as the model name, and price used to describe the general information of all products. Fig. 3 illustrates part of the electronic products ontology. Since the product features are specific for different product types, they are defined as attributes under each sub-class of product class. Fig. 4 illustrates the attributes in the scanner sub-class under the input class in Protege.

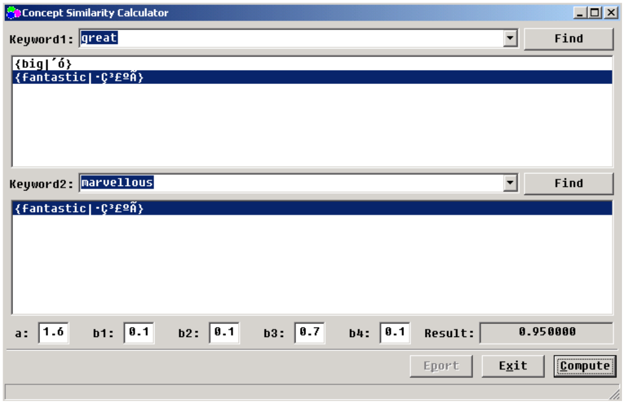


Fig. 1. Results of similarity between the two keywords "great" and "marvelous".

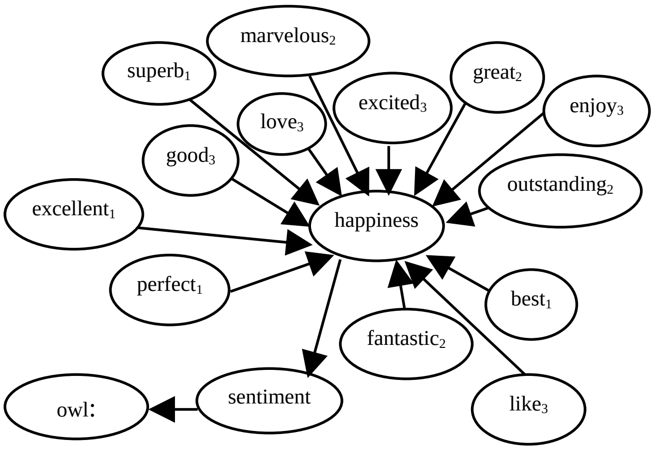


Fig. 2a. Emotions ontology: Happiness class with part of its sub-classes (including a subclass level number).

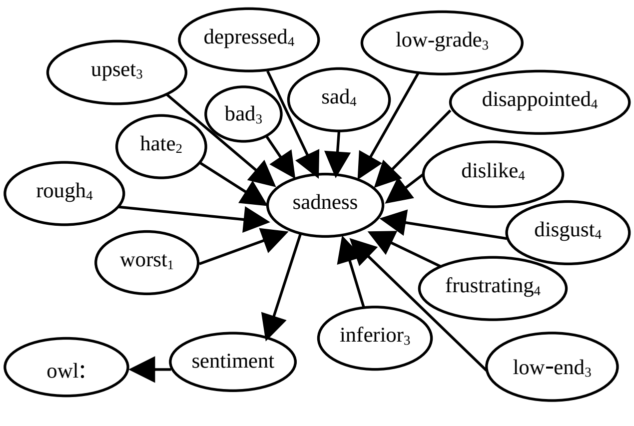


Fig. 2b. Emotions ontology: Sadness class with part of its sub-classes (including a subclass level number).

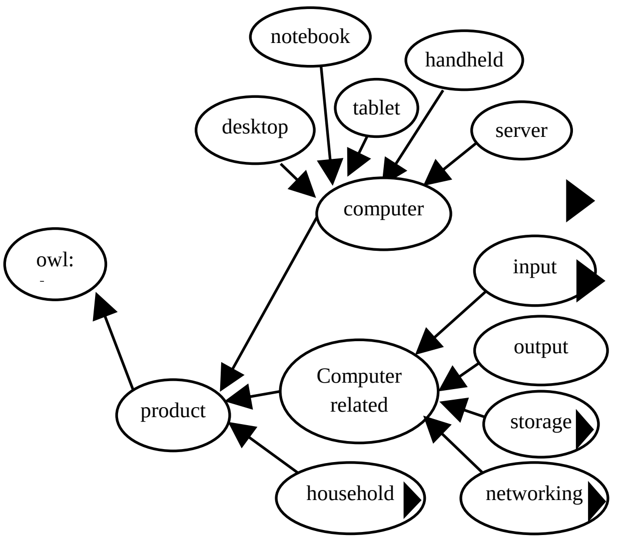


Fig. 3. Sample of the electronic products ontology with part of the sub-categories.

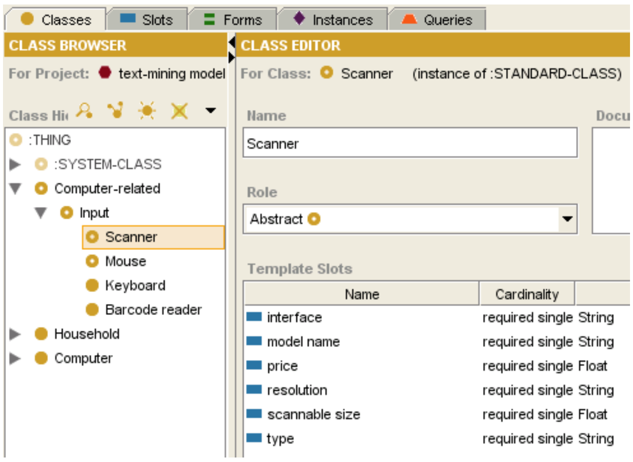


Fig. 4. Attributes in the scanner sub-class under the input class.

III. Methodology

The basic framework of the system is illustrated in Fig. 5, which includes four major processing modules: ontology management module, the user query processing module, information foundation module and query analysis engine module. All these different modules work together to complete the work of content retrieval according to the user's query. The functions of each part are described as follows.

A. *Ontology Management Module*

The ontology management module is responsible for the creation and maintenance of the electronic products ontology and the emotions ontology, which are the basis of the other functional modules.

B. *User Query Processing Module*

To demonstrate the analysis of a simple query, the following example is considered.

QUERY: "Which tablet PC is excellent?"

The main function of the user query processing module is segmenting and analyzing the query that users input, and then obtaining the different extracted words. The words in user queries can map directly to the concepts, attributes or ontology instances. We can use the corresponding concept, attribute or the instance to replace the user's query words. Keyword extraction is the process in which the query is processed for words that are infrequent and yield meaning to the entire query. First of all, parsing will be used to skip frequently used words and conjunctions such as: of, the, on, etc.; then, the remaining query words will be further analyzed by using a soft parsing technique which is applied to non-rule based data. For the purpose of the query considered above, the extracted keywords are *tablet* and *excellent*. The keyword *tablet* is extracted as it matches with *tablet*, which already exists in the electronic products ontology, whilst the keyword excellent is extracted due to its match with the emotions ontology.

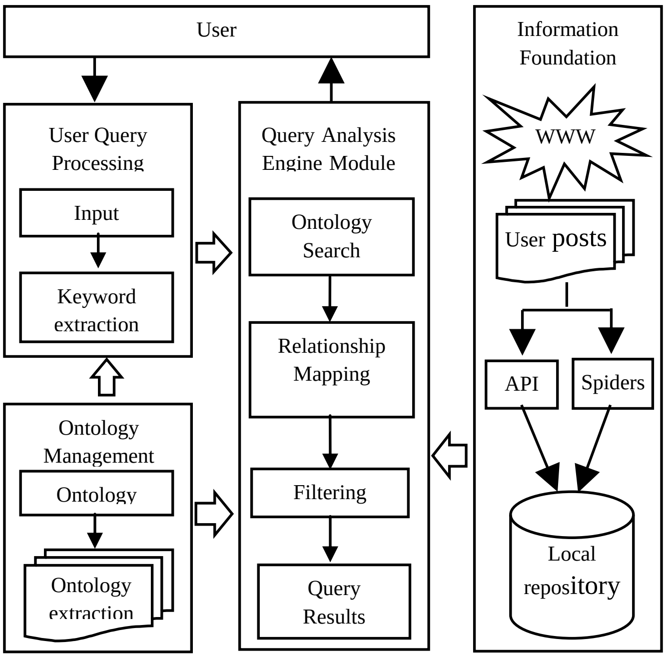


Fig. 5. Ontology-based sentiment analysis model for social networking websites.

The pseudo code for capturing the keywords is as follows:

TERMINOLOGY

Ke = KEYWORDS OF ELECTRONIC ITEMS

Ks = KEYWORDS OF SENTIMENT

x ÎC, if *x* is an entity that is an instance of class C

PSEUDOCODE

READ USER_QUERY

PARSE USER_QUERY

IF (WORD Î (SET OF ENTITIES IN ELECTRONIC ONTOLOGY))

K_e=WORD

ELSE IF (WORD Î(SET OF ENTITIES IN SENTIMENT ONTOLOGY))

K_s=WORD

Within the query or user posts, it is very common to have the negating terms (e.g. not and without) and words that enhance sentiment (e.g. love it very much and absolutely good). In this research, they will be taken into account. Based on the online survey of customer reviews on electronic products mentioned above, the examples of negating terms and enhancing terms are summarized in

Table I. Suppose A consists of all enhancing terms found in HowNet and B consists of negating terms, K_{qs} and K_{ps} are the emotional keywords found in a query q and a user post p respectively, $L_{K_{qs}}$ refers to the level at which the emotional keyword K_{qs} is located and $L_{K_{ps}}$ refers to the level at which the emotional keyword K_{ps} is located. The level of sentiment intensity is adjusted using the criteria shown in Fig. 6:

TABLE I: EXAMPLES OF NEGATING AND ENHANCING TERMS

Enhancing terms	Negating terms
Absolutely, very, so, lot, above, strongly, pretty, extremely	below, nothing , not, no, never, without, hardly, barely, rarely

Criteria 1: If a ∈ A comes with K_{qs}, $L_{K_{qs}}$ will be decreased by 1.
Criteria 2: If a ∈ A comes with K_{ps}, $L_{K_{ps}}$ will be decreased by 1.
Criteria 3: If b ∈ B comes with K_{qs}, $L_{K_{qs}}$ will be increased by 1.
Criteria 4: If b ∈ B comes with K_{ps}, $L_{K_{ps}}$ will be increased by 1.

Fig. 6. Different criteria showing the adjustment of sentiment intensity when negation and enhancing terms involved.

C. *Information Foundation Module*

A large pool of highly unstructured non-rule based data is available on social networking websites. User pages/posts are retrieved using the API provided by the social networking website or using spiders as explained in [2] and is stored in large text repositories locally.

D. *Query Analysis Engine Module*

This module consists of three parts: 1) Ontology search, 2) Relationship mapping and 3) Filtering. The output is the search results that the users really want. The pseudo code for the query analysis process is shown below:

TERMINOLOGY

K=SET OF KEYWORDS, K={*j*}
x ÎC, if *x* is an entity that is an instance of class C
S ÌC, if S is a subclass of C in the ontology
S:*P*=*V*, if *P* is the property of the subclass *S* and *V* is the set of values of *P*
ID(*x*, *y*), if a post consists of a join between keywords *x* and y

PSEUDOCODE

IF (*x* ÎK), SET OF RELATED WORDS = {*y* Î*x* | *y* ÌC *x* ÌC | *y*: (*y* ÎV *x*=P)}

LIST OF RESULTS= {ID(*x*, *y*)}

1) *Ontology search*

The extracted keywords are mapped to an extensive ontology that is based on the relationship of other concepts related to the domain. The keywords extracted in the above step will be mapped to the ontology as follows.

- *tablet* is a value of the *genre* property of the class *computer*. The entities having *genre: tablet* are instances of computer under the tablet category.
- *computer* is a generic class defining several entities.
- *excellent* belongs to the subclass *happiness* under the class *sentiment.*

2) *Relationship mapping*

The extracted keywords are mapped to an extensive ontology that is based on the relationship of other concepts related to the domain. The query engine establishes relationships between concepts similar to the keyword entity to give an extended meaning to the search. This is done by generating joins between the keywords and checking for existence of posts containing the keywords of the join. The results thus generated carry pages not just based on the keywords of the query but concepts surrounding them as well. In the case of our example, let us consider a pool of the following user posts/pages.

- **Post ID NO. 26**: *I love IPad2. It is very convenient for me.*
- **Post ID NO. 33**: *The new model of IPad2 is very cool.*
- **Post ID NO. 58**: *Lenovo ThinkPad X210 is good for beginners.*

In all the above cases, the system looks up words related to the keyword *excellent* within user posts containing instances of computer having the value of the genre property as *tablet* such as *IPad2* and *Lenovo ThinkPad X210.*

3) *Filtering*

The search results also contain ambiguous or redundant posts that are not relevant to the keyword. These results are filtered out, on the basis of their conceptual closeness to the keywords keeping only a fixed number of pages to be displayed as search results. In our example, the posts are filtered based on the following two stages:

a) Results that do not satisfy joins: Lenovo ThinkPad T410 is a great tablet PC

Although *great* is similar to the keyword *excellent, Lenovo ThinkPad T410* does not have its *genre* property set as *tablet.* Therefore, the post will be eliminated from the search results at this stage.

b) Results that consist of posts with average emotional level difference (ALD) of higher than the emotional tolerance index are eliminated

Inside a user post, there may be more than one emotional keyword. By mapping an emotional keyword of the user post and the query to the emotion ontology, the difference of subclass levels at which the two emotional keywords are located in the emotion ontology indicates their emotional difference. The average emotional level difference (ALD) is the average emotional difference between all emotional keywords of a user post and the emotional keyword of query. Suppose n is the emotional tolerance index indicating the level of tolerance to accept the emotional difference between query and user posts. The higher the value n, the larger the tolerance to accept the emotional difference between query and user posts since more user posts are included in the query result. It can be used as an adjustment parameter to restrict whether a user post p should be filtered by comparing the emotional difference between query and post p with the emotional tolerance index. The query result can be evaluated by referring to the emotion ontology based on the following algorithm:

Input: the user's query Q and the user post set $\{D\}$ which will be further filtered after stage 1 of filtering process.

Output: the user posts filtered as query result set $\{R\}$.
PROC TF$(Q,\ D_p)$
{
For each user post $D_p \in \{D\}$ do
{
For each emotional keyword in D_p and Q
{
count emotional keywords in D_p as Num_p
if (K_{qs} Ì A and K_{ps} Ì A) where A = {Happiness|Sadness}
TLD_p = TLD_p + Absolute ($L_{K_{qs}}$ - $L_{K_{ps}}$)
End if
}
ALD_p= TLD_p / Num_p
if $ALD_p \leq n$
$D_p \in \{R\}$
}
}

IV. Experimental Results

To evaluate our proposed model, four different query examples are used to get the results from 347 randomly selected customer reviews on facebook.com, which are fed into our proposed model for analysis. Query 1 involves the emotional keyword related to happiness while query 2 involves the emotional keyword related to sadness. Query 3 involves a negating term while query 4 involves an enhancing term.

Query 1: Which tablet PC is fantastic?
Query 2: Which tablet PC is the worst?
Query 3: Which tablet PC is not the best?
Query 4: Which tablet PC is very good?

Suppose the emotional tolerance index n = 0, 1 and 2, the query results of each query are compared with the manually evaluated results to find out the accuracy rate, which is shown in Table II. Negating and enhancing terms are also considered in the proposed model. Hence, suppose the emotional tolerance index n = 0, the accuracy rates of our proposed model are compared with the accuracy rate of the model without considering negating and enhancing terms shown in Table III.

TABLE II: Comparison of Accuracy Rates among Different Emotional Tolerance Indexes

	Emotional tolerance index n = 0	n = 1	n = 2
Q. 1	Accuracy – 97.2%	Accuracy – 62.4%	Accuracy –48.5%
Q. 2	Accuracy – 95.5%	Accuracy – 63.6%	Accuracy – 44.7%
Q. 3	Accuracy – 95.8%	Accuracy – 61.7%	Accuracy – 41.2%
Q. 4	Accuracy – 96.2%	Accuracy – 63.2%	Accuracy – 42.6%

TABLE III: Comparison of Accuracy Rates Between Proposed Model and the One Without Considering Negating and Enhancing Terms

	Our proposed model n = 0	Without considering negating and enhancing terms
Q. 1	Accuracy –97.2%	Accuracy – 63.2%
Q. 2	Accuracy – 95.8%	Accuracy – 75.9%
Q. 3	Accuracy –95.8%	Accuracy – 0.0%
Q. 4	Accuracy –96.2%	Accuracy – 0.0%

Based on the above experimental results, the accuracy rate exceeds 90% for all the four queries under emotional tolerance index 0. As the emotional tolerance index increases, the accuracy rate decreases rapidly for all the four different queries. Negating and enhancing terms are very important when sentiment analysis is performed.

V. Conclusion

With the advent of the Internet, social technology has become very important for online businesses as the social media technologies allow online businesses to analyze consumers' reviews of products. In this research, an ontology-based model is designed to analyze the sentiment of consumers of electronic products. The data is mined from a frequently updated source i.e. posts in the case of a social networking site. Therefore a system that updates associations between entities based on these additions is desirable. By using a query analysis engine that maps relations to the ontology; the system has the potential to display better semantic alignment capabilities. In addition, this proposed model takes the negating and enhancing terms into account in order to evaluate the sentiment level accurately. On the other hand, semantically, the relationships between keywords are dependent largely on the contexts in which they are used. Hence the results displayed, in certain cases, could be inaccurate as entities can be related through multiple contexts and links. Moreover, the language used on social networking websites is, to a large extent, non-rule based, i.e. it does not follow the syntactical rules of the English language. This requires the development of a highly detailed ontology, which is quite challenging. For such a complex system, it is difficult to predict its performance for different load conditions; this warrants a significantly larger study that will be the topic of future research. Although the electronic products domain is selected, it is possible to extend the project domain to various other relevant topics by building the respective ontologies. Different algorithms can be utilized to improve performance depending on the parameters that are of utmost importance. The accuracy rates of the proposed model are all over 90% with emotional tolerance index 0, which indicates little tolerance to accept emotional difference between query and user posts. Negating and enhancing terms have to be considered for sentiment analysis as they can greatly affect the accuracy.

References

[1] T. Mine, S. Lu, and M. Amamiya, "Discovering relationships between topics of conferences by filtering, extracting and clustering," in *Proc. 13th international Workshop on Database and Expert Systems Applications*, Washington, D.C, pp. 205-209, 2002.

[2] M. Castellano, G. Mastronardi, A. Aprile, and G. Tarricone, "A web text mining flexible architecture," *International Journal of Computer Science and Engineering*, vol. 1, no. 4, pp. 252–259, 2007.

[3] S. Pierre and V. Blondel, "Automatic discovery of similar words," *Survey of Text Mining: Clustering, Classification, and Retrieval Springer*, pp. 25-44, 2008.

[4] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proc. 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, San Jose, California, August, 2007.

[5] N. F. Noy and D. L. McGuinness, "Ontology development 101: a guide to creating your first ontology," Stanford University, USA, 2001.

[6] Protégé home page. [Online]. Available: http://protege.stanford.edu/

[7] H. Bast, F. Suchanek and I. Weber, “Semantic full-text search with ESTER: scalable, easy, fast,” *IEEE International Conference on Data Mining Workshops*, pp. 959-962, 2008.

[8] M. Enkhsaikhan, W. Wong, W. Liu, and M. Reynolds, “Measuring data-driven ontology changes using text mining,” in *Proc. the Sixth Australasian Conference on Data Mining and Analytics,* Gold Coast, Australia, December, 2007.

[9] K. M. Sam and C. R. Chatwin, “The mapping between business e-marketing mix and internet consumers’ decision-making styles in E-commerce,” in *Proc. the Fifth International Conference on Electronic Business*, Hong Kong, 2005.

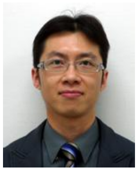
[10] K. M. Sam and C. R. Chatwin, “Ontology-based shopping agent for E-marketing,” *International Journal of Intelligent Information Technologies*, vol. 6, no. 2, pp. 21 - 43, 2010.

[11] H. Y. Gao and J. H. Zhao, “Ontology-based enterprise content retrieval method,” *Journal of Computers*, vol. 5, no. 2, pp. 314 – 321, 2010.

[12] C. W. Zhao, X. L. Yao, and S. H. Sun, “A hownet-based feature selection method for chinese text representation,” in *Proc. the sixth International conference on Fuzzy Systems and Knowledge Discovery*, Tianjin, China, 2009.

[13] Z. D. Dong and Q. Dong. HowNet. [Online]. Available: http://www.keenage.com

[14] L. Z. Liu, M. Y. Lei, and H. S. Wang, “Combining domain-specific sentiment lexicon with hownet for Chinese sentiment analysis,” *Journal of Computers*, vol. 8, no. 4, pp. 878 – 883, 2013.



Kin Meng Sam completed his PhD in Engineering and Design from the University of Sussex, United Kingdom in 2008.

He is currently an assistant professor in Electronic Business at the University of Macau. In 2010, he was elected as a member in Graduate Studies Committee. He is now a member in the Academic Council Committee at the University of Macau. He has published a few conference and journal papers regarding to online consumers’ decision-making styles, e-marketing mix and ontology. His current interests include i) e-marketing mix, ii) ontology-based business system, iii) online consumer behavior and iv) text-mining system. His current research is focused on designing an e-marketing mix analysis model. The aim of the research is to provide analyzed results to online businesses so that they can evaluate their performance of e-marketing mix in the market.

Dr. Sam is currently a member in IEEE (Institute of Electrical and Electronics Engineers) and AIS (Association for Information Systems).

Chris Chatwin is a professor in the Department of Engineering and Design, School of Engineering and Informatics, University of Sussex, United Kingdom.

He is currently holding the Chair of Engineering, University of Sussex, UK; where he is Research Director of the “iims Research Centre”. He has been a member of the University: Senate, Council and Court. He has published two research monographs: one on numerical methods, the other on hybrid optical/digital computing - and more than two hundred international papers. His current interests include i) recognition, tracking, surveillance & security systems, ii) informatics, communications & space, iii) biomedical diagnostics & prognostics and iv) e-commerce.

Prof. Chatwin is on the editorial board of the International Journal “Lasers in Engineering”. He is also a member of: IEEE (Institute of Electrical and Electronic Engineers); IEEE Computer Society; the British Computer Society; the Association of Industrial Laser Users, European Optical Society. He is a Chartered Engineer, Euro-Engineer, International Professional Engineer, Chartered Physicist, Chartered Scientist and a Fellow of: The Institution of Electrical Engineers, The Institution of Mechanical Engineers, The Institute of Physics and The Royal Society for Arts, Manufacture and Commerce.