A Reference Comments Crawler for Assisting Research Paper Writing

Hocheol Jeon
Agency for Defense Development, Korea

Abstract: *When writing a research paper, significant effort is spent comparing the current work to other related studies. In general, these comparisons comprise the 'Related Work' section of the paper, with the relevant reference papers cited and analyzed. An automatic method for gathering and managing information about other researchers' reviews of reference papers would reduce the time and effort needed for such comparisons. Thus, in this paper, we propose a crawler that gathers the comments made by other researchers on the papers that are cited in the 'Related Work' section and listed in the 'Bibliography' or 'References' section of a research paper. The Reference Comments Crawler (RCC) system collects the text pertaining to the reference papers, providing useful information to researchers by extracting relevant data from the comments. The RCC considers different types of reference identifiers and the comment sentences are extracted based on these reference identifiers and user-defined extraction rule. Also, the RCC system extracts and provides the previous and subsequent sentences, labeled as PreSentences and PostSentences, as well as the comment sentences including the reference identifier. A series of experiments were performed to evaluate precision and recall, and the results showed that the RCC system can provide useful information with a high degree of precision and recall to the user. Furthermore, through these experiments, our system can assist researchers by reducing the time and effort spent comparing and analyzing related work.*

Keywords: *RCC, comment sentences, reference identifiers, research paper writing.*

Received March 13, 2012; accepted April 11, 2013; published online February 26, 2014

1. Introduction

Researchers write technical papers to publicize their research accomplishments in their own professional fields. It is a globally-recognized convention for most engineering research papers to have a 'Related Work' section, in which an analysis of related studies is found, comparing and contrasting previous work with the current work. The papers cited in the 'Related Work' section are also, listed in the 'References' section.

In general, a citation about a reference paper consists of a comment sentence and the sentences surrounding it that we refer as preSentences and postSentences. We define a comment sentence as a sentence in a research paper that describes and evaluates the main ideas of a cited reference paper. Typically, a comment sentence describes notable features of the paper and pinpoints the pros and cons of its methodology. A comment sentence naturally contains a reference identifier that uniquely identifies the corresponding reference paper. A comment sentence with a reference identifier is usually referred to as a citation.

We analyzed 110 journals that were included in the 'Computer Science, Information'category of the Journal Citation Reports (JCR), (http://admin-apps. isiknowledge.com/JCR/JCR?PointOfEntry=Home) to determine the feasibility of our research. According to this analysis for the year 2009, the total number of published papers was 7,455, with 212,965 references, for an average of 32 references per paper.

In general, analyzing the related publications is necessary. This task can make paper writing more difficult, time-consuming, and tedious, as a large portion of the researchers' time and effort is spent comparing their current work with other related studies. Thus, many authors attempt to determine the key points and characteristics of each study as quickly as possible. If there were a fast and accurate way to analyze the related literature, researchers could focus more on the main content of their papers and have more time to spend on performing precise and frequent experiments, while retaining the quality of the research and subsequent write-up.

In this research, we propose a crawler that functions by collecting the comments made by other researchers on the papers that are cited in the content of the research paper. All comments are valuable and important because they summarize the essential parts of the reference papers and describe their pros and cons. Moreover, comments about the same paper are often expressed with opposing opinions due to differing viewpoints, providing a variety of different perspectives for reference papers.

We implemented the Reference Comments Crawler (RCC) system that collects comments pertaining to the reference papers and extracts relevant data from the comments. The operation of the RCC is largely dependent on a reference identifier.

For our evaluation, the RCC gathered reference papers cited in randomly selected research papers;

however, the papers with limited access were collected manually. A series of experiments were performed to evaluate the performance of the RCC system, and the results showed that it provided relevant comment sentences with relatively high precision and recall.

2. Related Work

2.1. Opinion Mining

Opinion mining is a branch of data mining [24] that analyzes individual subjective opinions such as product reviews and extracts meaningful information from these reviews, including the orientation of opinions [1, 4]. Opinion mining has been applied in many domains including products [1, 4, 5, 12], movies [27, 28], music [6, 8, 19], hotel [13], and news [2, 14].

For product reviews, opinion mining extracts product features such as color, battery life, and price [1]. Some studies define linguistic rules to extract features [4], and other studies employ a holistic lexicon-based approach using external evidence and linguistic conventions to identify the semantic orientations of opinions [5]. The accuracy of identification increases using Part-Of-Speech (POS) information, which indicates the relationships among the extracted features [12].

For movie reviews, most studies focus on analyzing and summarizing user reviews [28], then manually building a movie ontology and marking its nodes as positive, negative, or neutral using opinion sense scores [27].

For music reviews, most studies determined appropriate times of play for a song by detecting usage terms such as "I would listen to this song while driving, sleeping, etc." through user review analysis [8], and other studies analyzed online music reviews to identify terms people use when commenting negatively or positively [6].

Opinion mining for news reviews analyzes comments on news articles and provides readers with news articles of related content. A related study utilized FrameNet, an online lexical database with 800 semantic frames, to identify opinions and opinion-holders [14]. Another study classified each user comment as positive or negative by counting the numbers of stronger and weaker positive and negative words in articles [2].

The opinions gathered are classified as positive or negative according to an ontology [15, 27] or machine learning methods, such as the Hidden Markov Model (HMM) [11]. Recently, opinion mining methods have been applied to blogs [20, 23] and Twitter [18, 21], in addition to the web

While most studies of opinion mining focus on utilizing customer reviews available on the web, blogs, and Twitter to provide useful information to sellers or manufacturers, our study considers research papers available on the web and analyzes only the comment sentences, rather than entire papers.

2.2. Focused Crawling

Focused crawling attempts to download only documents about a particular topic. Hence, each collection gathered by a focused crawler will be much smaller than that gathered by a general web crawler. However, it is only effectively used for certain fields, as it searches only the subset of the web related to a specific topic, and offers potential solutions to the vertical search problem [17, 22].

Focused crawlers are used for various purposes: to enrich vocabularies for ontology concepts [17], to design ontologies of personal web pages or construct ontology classes of such pages [25], and to collect related e-commerce information using an e-commerce ontology [9]. It can also, improve the quality of web navigation by utilizing content and link structure analyses [3, 10, 22], employing various learning methods [7, 26], or using a domain ontology [16].

Most studies of focused crawling emphasize the gathering of all available documents related to a particular topic. Our proposed system, however, only detects research papers that are cited by a particular paper.

3. The Reference Comments Crawler System

We implemented the RCC system to collect comments about reference papers and provide useful information by extracting relevant data. The RCC has two main functions: Reference paper crawling and comment sentence extraction.

Reference paper crawling finds and downloads papers that include citations of reference papers included in the bibliography of a selected paper by searching Google Scholar (http://scholar.google.com/) using the title of the reference paper. The reference papers themselves are not downloaded, as they do not contain the comments. Instead, papers containing comments about the pertinent reference papers are collected.

Comment sentence extraction finds and selects sentences that mention reference papers. The extraction is accomplished based on the reference identifier that is used to identify the cited reference papers. The reference identifier can be composed of digits, author names, or the combination of author names and publication years.

To facilitate these functions, the RCC system has a system architecture that consists of the Crawler, the Filter, the Analyzer, and the User Interface, as shown in Figure 1.

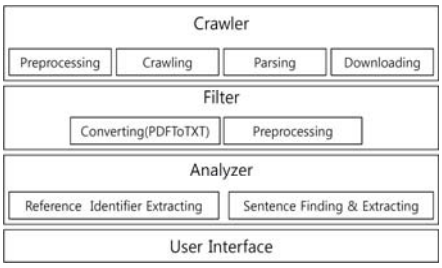


Figure 1. The system architecture of the RCC.

The User Interface interacts with the user, receiving a paper title during the reference paper crawling phase. During the comment sentence extraction phase, it receives a directory path, displaying the comment sentence and some of the surrounding sentences by identifying a sentence from a PDF file in the directory. The Crawler searches for papers that include reference paper citations, creates a list of the papers obtained from Google Scholar, extracts their URLs, downloads the PDFs, and stores them in an automatically created directory. The Filter converts each PDF file to TXT format to facilitate the extraction of comment sentences. The Analyzer finds reference identifiers from the collected papers and extracts comment sentences based on these identifiers.

4. Reference Paper Crawling

Reference paper crawling searches for the papers that have citations about the reference papers listed in the 'References' section of the current paper. To accomplish this, the title of a reference paper is queried to Google Scholar and the papers found are downloaded. Note again that we do not download the reference papers themselves since they do not contain the comments about their papers; instead we download the papers that pertain the comments about the reference papers.

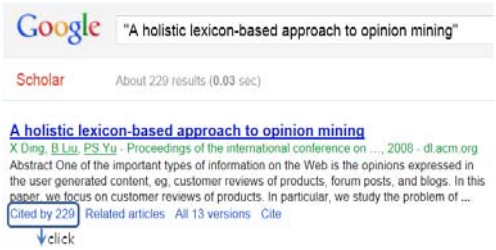
Reference paper crawling starts by inputting the title of a reference paper to the user interface of the RCC, which launches the Crawler and begins a Google Scholar search. Then, the RCC Parser analyzes the search results and finds papers that contain citations for the reference papers. Finally, the RCC Downloader downloads the corresponding PDF files and stores them in a system directory.

As an example, suppose you want to find papers that contain citations about a reference paper titled "A holistic lexicon-based approach to opinion mining". The first step is to query the given paper title to Google Scholar as shown in Figure 2-a. In general, the topmost result would be the paper exactly matched with the query, as is the case in the Figure. Note that Google Scholar provides information about how many papers cite the target reference paper at the bottom of the search result display, as depicted in Figure 2-a expressed by a link with anchor text "Cited by 119".

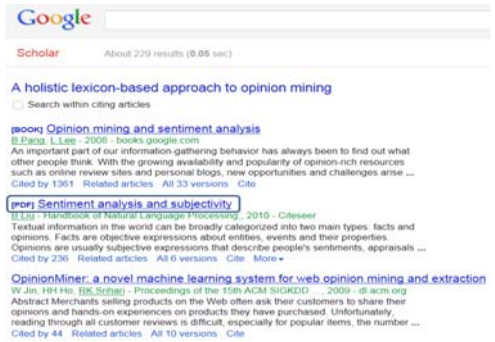
In the next step, by clicking this "Cited by 119" link, we obtain a list of papers that include the citations about the target reference paper, as shown in Figure 2-b. Among these papers, we select those that provide corresponding PDF files, as indicated by a box in the figure, and download and store them in a system directory.

Figure 2-c is given to confirm that the selected paper in the list in fact contains citation and reference list for the target reference paper.

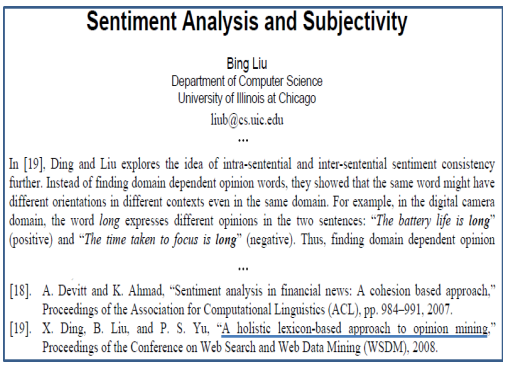
Reference paper crawling is somewhat restrictive, because many papers often have limited access; therefore, it is not possible to collect all of the possible searchable papers through automatic gathering. Hence, in addition to automatic paper crawling, Our RCC system has the option of gathering papers manually.



a) Querying the title of a reference paper to Google Scholar.



b) A list of papers that include citations about the reference paper.



c) An example paper that contains citation and reference list for the reference paper.

Figure 2. An example of reference paper crawling.

5. Comment Sentence Extraction

5.1. Reference Identifier Extraction

A reference identifier is used to identify each reference paper and is expressed in various forms. Figure 4 shows five examples of reference paper descriptions. Each example has a different form of reference identifier, including: a digits-only number (e. g., 19), abbreviated author names with a publication year (e. g., [Wang *et al.* 2003]), full author names with a publication year (e. g., Wooldrige M. and Jennings N., 1995), a letter-digit word (e. g., [ZP05]), and a number with a bracket (e. g., [41]).


19. R. Want, K. P. Fishkin, A. Gujar, and B. L. Harrison. "Bridging Physical and Virtual Worlds with Electronic Tags". Proceedings of SIGCHI '99 (Pittsburgh, PA, May 15-20) ACM, New York, 1999. pp. 370-377

[Want et al., 2003] Want, R., Pering, T., and Tennenhouse, D. (2003). Comparing autonomic and proactive computing. *IBM Systems Journal in Technology*, 42(1):129–135.

WOOLDRIDGE, M. AND JENNINGS, N. 1995. Intelligent agents: Theory and practice. *Knowledge Engineering Review 10*, 2, 115–152.

[ZP05] H. Zheng, and C. Peng, "Collaboration and Fairness in Opportunistic Spectrum Access," In *Proc. 40th annual IEEE International Conference on Communications*, Jun 2005.

[41]. Sparacino F. 'Museum Intelligence: Using Interactive Technologies For Effective Communication And Storytelling In The Puccini Set Designer Exhibit.' In: *Proceedings of ICHIM 2004*, Berlin, Germany, August 31-September 2nd 2004.

Figure 3. Examples of reference identifier expressions.

Reference identifier extraction is important in the RCC, as it serves to identify comment sentences som, that the reference identifiers contained in the citation can be matched with those in the corresponding description in the reference list. Thus, it provides a connection between the comments and their corresponding reference papers. We designed and developed an algorithm for reference identifier extraction and its pseudo code is presented in Algorithm 1. It is implemented as a function with two arguments; the title of a reference paper and the content of a crawled paper.

Algorithm 1: Reference Identifier Extraction

1. *function RidExtraction(title, content) // preprocessing the content of a paper to get 'References' section of the paper*
2. *ref_content ← getReferencesSection(content); // find the location of title in ref_content*
3. *titleLoc ← findLoc(title, ref_content); // find out the type of Rid (digit-only or begins with a letter)*
4. *Boolean isDigit ← getRidType(ref_content); // determine whether a bracket surrounds Rid*
5. *Boolean hasBracket←getBracketforRid(ref_content);*
6. *if hasBracket then*
7. *openBracketLoc←findLastOpenBracketLoc(titleLoc, ref_content);*
8. *closedBracketLoc←findClosedBracketLoc(openBracketLoc, ref_content);*
9. *Rid ← extractRid(openBracketLoc, closedBracketLoc, ref_content);*
10. *else if isDigit then // Rid is digits('D') only*
11. *Rid ← getNextRid('D', titleLoc, ref_content);*
12. *else // Rid begins with a letter ('L')*
13. *Rid ← getNextRid('L', titleLoc, ref_content);*
14. *end if*
15. *return Rid;*
16. *end function*

The algorithm starts by preprocessing the content of a paper, completing tasks such as removing tabs, and gets the 'References' section of the content (Line 3). Then, the title of the given reference paper is searched through this reference content to find the location of the paper in the list (Line 5). At this point, we know that the reference identifier of the target paper is positioned at the beginning of this location.

We consider three situations relating to the forms of reference identifiers in particular. The first situation occurs when a square bracket is used in the reference identifier expression. It is relatively easy to find the reference identifier in this case because it is necessary to find the closest open and closed square brackets (Lines 10-13). Square brackets should be validated because they are widely used for other purposes, such as range representation.

The second situation is when a reference identifier is expressed using a digits-only number.

In this case, a reference identifier can be detected using a search template "\n digit+ .", representing that a number consisting of digits occurring between a newline('\n') and a period('.') is recognized as a reference identifier (Lines 14-15). (Note that '+' symbol in the template indicates that digit occurs at least one time.) Validation should also, be carried out after template matching based on the following conditions; A blank or a newline has to occur after template matching., an author's name, beginning with a capital letter, must follow a blank or a new line.

The last situation occurs when a reference identifier begins with a letter and is expressed using letters and digits. A search template for this case would be "\n letter (letter | digit)* ." (Lines 16-17). (Note that '|' symbol in the template indicates optional appearance and '*****' symbol indicates repetition of zero or more times.) Again, validation checking should be performed based on the following conditions; Every reference ends with a period except for URL., two references are separated by a newline.

5.2. Comment Sentence Extraction

Comment sentence extraction begins with directory selection, after which PDF files in the selected directory are sequentially loaded by the Loader and converted to TXT-formatted files by the Filter. After preprocessing, the RCC Analyzer finds the reference papers in the 'References' section of these converted files using the given paper title and extracts their associated reference identifiers. This is accomplished by the reference identifier extraction algorithm described in the previous section. Finally, it finds and extracts the comment sentences from the content of

the current paper using these reference identifiers. During this process, in addition to a comment sentence, the previous and subsequent sentences are also, extracted, since there are many cases in which the sentences surrounding the comment sentence provide useful information about the reference paper.

We developed an algorithm for comment sentence extraction that the reference identifier extraction algorithm, and a pseudo code is given in Algorithm 2.

Algorithm 2: Comment Sentence Extraction

1. *procedure CommentExtraction(title, content) // extract Rid using the RidExtraction function*
2. *Rid ← RidExtraction(title, content); // preprocessing the content of the paper to remove 'References' section*
3. *main_content ← removeReferences(content); // find Rid in the content of the paper*
4. *ridLoc ← findNextRidLoc(Rid, main_content);*
5. *while ridLoc is valid do // repeats for all Rids*
6. *sentence ← extractSentence(ridLoc, main_content);*
7. *preSentences[]←extractPreSentences(ridLoc, main_content);*
8. *postSentences[]←extractPostSentences(ridLoc, main_content);*
9. *comments[i]←createComment(Rid, sentence, preSentences, postSentences); // searches for more Rid appearances*
10. *ridLoc ← findNextRidLoc(Rid, main_content);*
11. *end while*
12. *end procedure*

The algorithm first finds and extracts the reference identifier using the reference identifier extraction algorithm (Line 3), and then simply searches for sentences that contain it (Line 9), which are then identified as comment sentences. The RCC also, extracts the previous and subsequent sentences, labeled as PreSentences and PostSentences (Lines 10-11). Together, these data - the reference identifier, comment sentence, PreSentences, and PostSentences - comprise a comment unit that is delivered to the user (Line 12). A comment unit can be defined as the following vector:

$$Comment=<Rid,commentSentence,PreSentences,PostSentence> \quad (1)$$

Once extracted, each comment unit is stored in the comment sentence array and, at the same time, displayed on the user interface as shown in Figure 4.

6. Results and Discussion

We evaluated the effectiveness of the RCC in terms of precision and recall measures.

We randomly selected 43 research papers. The total number of references cited in these papers was 6594, among which we selected 3291 references that were written in English. We further examined these papers and found that only 2583 of the papers included reference identifiers, while the remaining 708 papers contained errors and were thus removed. Among these, the number of papers including comment sentences was 2400, while the remaining 183 papers did not contain any comment sentences, although they may have cited reference papers.

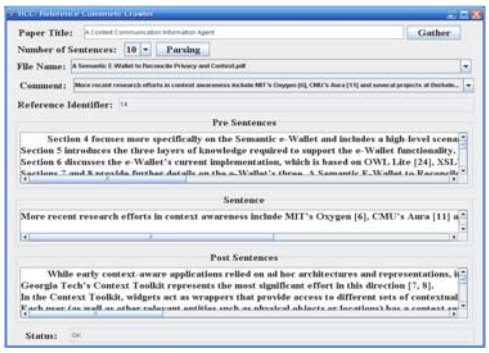


Figure 4. User Interface of the RCC.

Figure 4 shows the user interface of the RCC, displaying a comment sentence with PreSentences and PostSentences extracted by the comment sentence extraction module. Note that the 'Number of Sentences' option in the interface controls the sizes of PreSentences and PostSentences.

To evaluate the effectiveness of the RCC, we defined precision and recall as follows. Precision is the ratio of the number of correct comment sentences extracted over the total number of extracted comment sentences. Recall is the ratio of the number of correct comment sentences extracted over the total number of correct comment sentences.

Figures 5 and 6 show the precision and recall measures for the test collection, respectively. The average precision and recall were very high, 0.94 and 0.91, respectively.

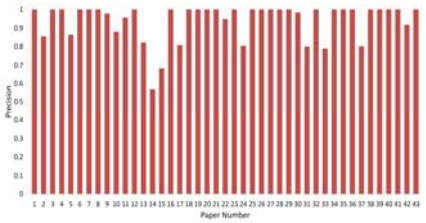


Figure 5. Precisions for the test papers.

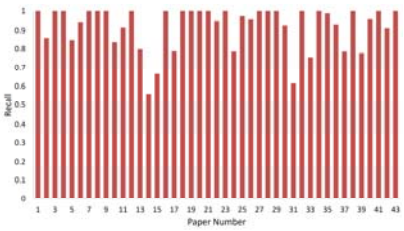


Figure 6. Recalls for the test papers.

There are few cases in which the precision of the RCC is relatively low, primarily because the description of a reference identifier in the 'References' section of the paper is different from the description of the same identifier in the body of the paper. When there are multiple reference papers for the same author in the 'References' section, each

citation in the content is identified by a reference identifier in which information about the publication year is added. As most reference identifiers consist of only digits or letters, there are only a few cases that include both the author name and the publication year.

The recall measures also, contained some relatively low values. This occurs when a reference identifier is specified as a superscript in the body of the paper, or when the extracted reference identifier is digit only, while the citation in the body of the paper is specified in terms of author names.

7. Conclusions

We proposed a novel tool to aid research paper writing, aimed at drastically reducing the time and effort spent comparing and analyzing related work. Our method utilizes other researchers'comments about the reference papers. A series of experiments verified that the RCC system performs effectively with a high degree of precision and recall.

We are planning to employ a variety of more appropriate rules to improve the accuracy of extracting comment sentences. In addition to extracting a whole sentence as a comment, we are devising a method of extracting the features and properties of a reference paper by analyzing comment sentences and their previous and subsequent sentences using opinion mining techniques.

References

[1] Archak N., Ghose A., and Ipeirotis P., "Show me the Money! Deriving the Pricing Power of Product Features by Mining Consumer Reviews," *in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, pp. 56 - 65, 2007.

[2] Balahur A., Steinberger R., Goot E., Pouliquen B., and Kabadjov M., "Opinion Mining on Newspaper Quotations," *in Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Milano, Italy, pp. 523 - 526, 2009.

[3] Chen X. and Zhang X., "HAWK: A Focused Crawler with Content and Link Analysis," *in Proceedings of IEEE International Conference on E-business Engineering*, Xian, China, pp. 677 - 680, 2008.

[4] Ding X., Liu B., and Yu P., "A Holistic Lexicon-Based Approach to Opinion Mining," *in Proceedings of the International Conference on Web Search and Web Data Mining*, Stanford, USA, pp. 231 - 240, 2008.

[5] Ding X., Liu B., and Zhang L., "Entity Discovery and Assignment for Opinion Mining Applications," *in Proceedings of the 15th ACM SIGKDD International Conference on Know-Ledge Discovery and Data Mining*, Paris, France, pp. 1125 - 1133, 2009.

[6] Downie J. and Hu X., "Review Mining for Music Digital Libraries: Phrase II," *in Proceedings of the 6th Joint Conference on Digital Libraries*, Chapel Hill, USA, pp. 196 - 197, 2006.

[7] Gao Z., Du Y., Yi L., Yang Y., and Peng Q., "Focused Web Crawling Based on Incremental Learning," *Journal of Computing Information Systems*, vol. 6, no. 1, pp. 9 - 16, 2010.

[8] Hu X., Downie J., and Ehmann A., "Exploiting Recommended Usage Metadata: Exploratory Analyses," *in Proceedings of the International Conference on Music Information Retrieval*, Victoria, Canada, pp. 19 - 22, 2006.

[9] Huang W., Zhang L., Zhang J., and Zhu M., "Semantic Focused Crawling for Retrieving E-Commerce Information," *Journal of Software*, vol. 4, no. 5, pp. 436 - 443. 2009.

[10] Jamali M., Sayyadi H., Hariri B., and Abolhassani H., "A Method for Focused Crawling Using Combination of Link Structure and Content Similarity," *in Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, Hong Kong, China, pp. 753 - 756, 2006.

[11] Jin W. and Ho H., "A Novel Lexicalized HMM-Based Learning Framework for Web Opinion Mining," *in Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 465 - 472, Montreal, Canada, 2009.

[12] Joshi M. and Penstein-Rosé C., "Generalizing Dependency Features for Opinion Mining," *in Proceedings of the 47th ACL and the 4th IJCNLP Conference Association for Computational Linguistics*, Singapore, pp. 313 - 316, 2009.

[13] Kasper W. and Vela M., "Monitoring and Summarisation of Hotel Reviews," *in Proceedings of International Conference on Information and Communication Technologies in Tourism*, Helsingborg, Sweden, pp. 471 - 482, 2012.

[14] Kim S. and Hovy E., "Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text," *in Proceedings of the Workshop on Sentiment and Subjectivity in Text*, Stroudsburg, USA, pp. 1 - 8, 2006.

[15] Lau R., Lai C., Ma J., and Li Y., "Automatic Domain Ontology Extraction for Context-Sensitive Opinion Mining," *in Proceedings of the 30th International Conference on Information Systems*, Phoenix, Arizona, pp. 35 - 53, 2009.

[16] Liu Z., Du Y., and Zhao Y., "Focused Crawler Based on Domain Ontology and FCA," *Journal*

of Information & Computation Science, vol. 8, no. 10, pp. 1909 - 1917, 2011.

[17] Luong H., Gauch S., and Wang Q., "Ontology-Based Focused Crawling," *in Proceedings of the International Conference on Information, Process, and Knowledge Management*, Mexico, pp. 123 - 128, 2009.

[18] Maynard D., Bontcheva K., and Rout D., "Challenges in Developing Opinion Mining Tools for Social Media," available at: http://gate.ac.uk/sale/lrec2012/ugc-workshop/opinion-mining-extended.pdf, last visited 2012.

[19] Meintanis K. and Shipman F., "Visual expression for organizing and accessing music collections in music-wiz," *in Proceedings of the 14th European Conference on Re-search and Advanced Technology for Digital Libraries*, Glasgow, UK, pp. 80 - 91, 2010.

[20] O'Hare N., Davy M., Bermingham A., Ferguson P., Sheridan P., Gurrin C., and Smeaton A., "Topic-Dependent Sentiment Analysis of Financial Blogs," *in Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*, Hong Kong, China, pp. 9 - 16, 2009.

[21] Pak A. and Paroubek P., "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in *Proceedings of the 7th Conference on International Language Resources and Evaluation*, Valletta, Malta, pp. 19 - 21, 2010.

[22] Pal A., Tomar D., and Shrivastava S., "Effective Focused Crawling Based on Content and Link Structure Analysis," *the International Journal of Computer Science and Information Security*, vol. 2, no. 1, pp. 1 - 5, 2009.

[23] Tsai T., Shih C., Peng T., and Chou S., "Explore the Possibility of Utilizing Blog Semantic Analysis for Domain Expert and Opinion Mining," *in Proceedings of the International Conference on Intelligent Networking and Collaborative Systems*, Barcelona, Spain, pp. 241 - 244, 2009.

[24] Yafi E., Al-Hegami A., Alam A., and Biswas R., "YAMI: Incremental Mining of Interesting Association Patterns," *the International Arab Journal of Information Technology*, vol. 9, no. 6, pp. 504 - 510, 2012.

[25] Yang S. and Hsu C., "Ontology-Supported Focused-Crawler for Specified Scholar's Webpages," *in Proceedings of IEEE the 8th International Conference on Intelligent Systems Design and Applications*, Kaohsiung, Taiwan, pp. 409 - 414, 2008.

[26] Yohanes B., Handoko, and Wardana H., "Focused Crawler Optimization Using Genetic Algorithm," *Journal of TELKOMNIKA*, vol. 9, no. 3, pp. 403 - 410, 2011.

[27] Zhao L. and Li C., "Ontology Based Opinion Mining for Movie Reviews," *in Proceedings of the 3rd International Conference on Knowledge Science, Engineering and Management*, Vienna, Austria, vol. 5914, pp. 204 - 214, 2009.

[28] Zhuang L., Jing F., and Zhu X., "Movie Review Mining and Summarization," *in Proceedings of the 15th ACM International Conference on Information and Knowledge Manage-Ment*, Arlington, USA, pp. 43 - 50, 2006.

Hocheol Jeon received his MS and PhD degrees in Computer Science and Engineering from Hanyang University, Korea, in 1999 and 2011, respectively. He is currently a researcher at Agency for Defense Development, Korea. His research interests include intelligent agents, information retrieval, information extraction, and context-aware user modelling.