Development of Framework System for Managing the Big Data from Scientific and Technological Text Archives

Mi-Nyeong Hwang[1], Myunggwon Hwang[1], Ha-Neul Yeom[1,4],
Kwang-Young Kim[2], Su-Mi Shin[3], Taehong Kim[1], and Hanmin Jung[1,4,]*

[1]Dept. of Computer Intelligence Research, [2]Dept. of Overseas Information,
[3]Dept. of NDSL Service, Korea Institute of Science and Technology Information, Korea
[4]Korea University of Science and Technology, Korea
{mnhwang,mgh,lucetesky,glorykim,sumi,kimtaehong,jhm}@kisti.re.kr

Abstract. In today's era of big data, increasing attention is being paid to the relationships among different types of data, and not just to those within one type of massive data, while processing and analyzing these data. To analyze and predict the trends of technologies from literatures on the basis of the conventional form of textual documents, such as academic papers or patents, the objects of analysis should include recent information collected from news websites and social media sites, which indicate the user preferences. It is necessary to systematically collect multiple texts to integrate and analyze different types of data. This study introduces practical ways to implement a database on the basis of the global standard using the unstructured information management architecture (UIMA).

Keywords: Big data, Text Big data, Scientific and Technological Text, Text Crawling System, Web Crawler, SNS Crawler, UIMA

1. Introduction

At the 2012 World Economic Forum Annual Meeting in Davos, Switzerland, the big data processing technology was highlighted as the "most important scientific technology of the year" [1]. According to IBM, 80% of big data are scattered and unstructured and hence, cannot be managed as structured data. To process these scattered data, we need to develop a new method of collecting and analyzing data [2]. Businesses were the first to understand the value of the analysis of unstructured data. This analysis has now been expanded from conventional data such as those from academic papers, patents, and magazines to data extracted from news websites and social network services such as Twitter and Facebook in order to build business intelligence [3]. Such an expanded application of data analysis is utilized to analyze and predict the trends of the advances of scientific technology [4].

A big data processing platform consists of the four steps of collection, storage, analysis, and visualization [5]. To maximize the application of analysis and visualization, a thorough collection and an appropriate storage of big data are needed.

* Corresponding Author.

This paper explains the collection of big data related to scientific technology from different sources; this process is necessary to analyze the trends of scientific technology. It is expected to help researchers to improve their research and to better implement a database.

2 Related Work

The introduction and penetration of the World Wide Web has led to an increasing effort to crawl data from documents on the Web [6]. As the size of the Web increases, more studies are being conducted on the collection of documents on certain topics from the Web than on their storage in one place [7,8]. There have been efforts to develop a crawling process of scattered data to collect large documents. In this case, there are certain disadvantages related to the sorting of overlapped data and to data storage and management [9]. Furthermore, there has been research on the geographical partition of the server storing the original documents, focusing on the speed of the scattered crawlers of mass data [10]. However, a more macroscopic approach is needed in this era of big data where multiple sources of unstructured data are scattered because these crawlers focus on optimizing the crawling performance for one type of data. This paper suggests a framework for collecting multiple texts on scientific technology.

3 Scientific Technology Text Crawling System

Figure 1 shows the process of implementing the data collection system for texts related to scientific technology. Unstructured data related to scientific technology from academic papers, patents, Wikipedia, news websites, and social media were first collected as raw HTML, XML, and text data.

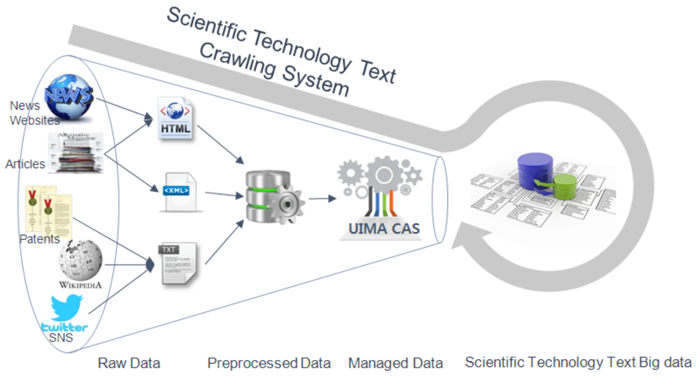

Fig. 1. Flow of scientific technology text crawling system

Metadata, which are needed to analyze papers related to scientific technology, were extracted from the collected data through this preprocessing procedure. The metadata were transformed into the common analysis structure (CAS) format of the unstructured information management architecture (UIMA) and were processed for implementing the data of big data texts on the basis of the global standard.

3.1 News Websites

News websites are an important source for collecting the latest news on scientific technology considering the fact that there is a gap between the time of research and the publication of academic papers and patents. 154 websites, whose services include news, magazines, and forums, such as Scientific Computing[1], ScienceNews[2], and Bioscience[3], were selected to collect the latest news on scientific technology. Data from news websites were collected using three types of crawlers, namely Google crawlers, RSS crawlers, and direct crawler, as shown in Figure 2, because they contain news published since 2001.

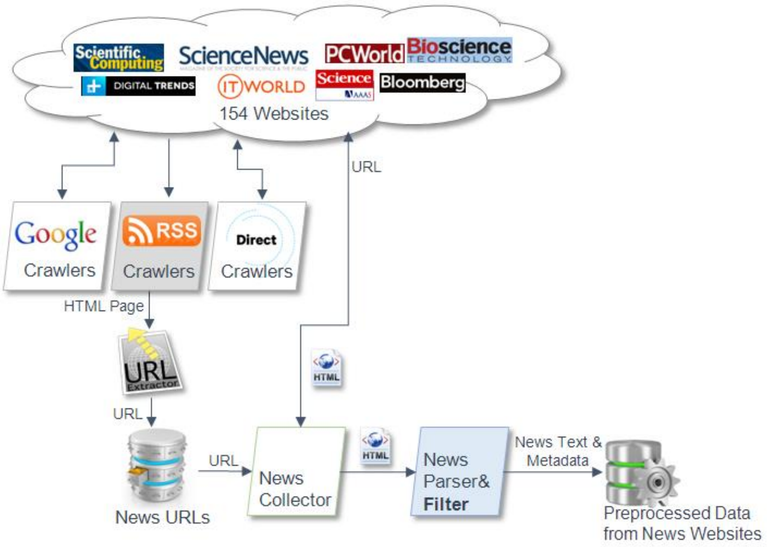

Fig. 2. Process of crawling news websites

News articles, which were already published in 2001, were collected by using both Google crawlers and direct crawlers that collect data from the websites. In the case of a website that is run by a keyword-based search engine, the direct extraction process showed the search results when keywords related to scientific technology were used. From a website that shows lists of news articles, the crawlers extracted the links of the news. From websites that provide an RSS service, the crawlers collected links of real-

[1] http://www.scientificcomputing.com/
[2] http://www.sciencenews.org/
[3] http://www.biosciencetechnology.com/

time news in a parallel manner. These collected URLs were stored in the URL database to avoid overlapping data. Then, the News Collector collected the HTML code of the news items through the URLs, and the News Parser extracted the title, author, date, category, and the content from the HTML code. The News Filter eliminated the overlapping and irrelevant news.

3.2 Articles

In the case of collecting foreign papers from websites such as IEEE and NCBI PubMed, data were directly collected from the websites by using the news website crawler. Meta information, which includes the title, author information, keywords, and abstract of a paper, was collected in this manner. The objects of the real-time data collection included data published between 2001 and 2014.

3.3 Patents

Patent data are also needed to analyze the levels of originality and scientific progress. Patent data released internationally, particularly in the US and Europe, and registered in the US between 2001 and 2013 were collected in bulk. The metadata and abstracts from these data were used in this study.

3.4 Wikipedia

Wikipedia[1] is an Internet encyclopedia, whose contents are created directly by the users. Here, a uniform resource identifier (URI) is assigned to every piece of information and DBpedia[2], which provides the related meta information, is downloaded to implement the database.

3.5 Social Network Service Data

Along with data from papers, patents, news websites, and Wikipedia, data from social network services were collected. Among the social media contents, we collected tweets. Tweets that included 213 keywords on scientific technology, such as web, computer, and smartphone, were collected real-time using OpenAPI released by Twitter[3]. Data from 2014 were collected, and on average, about 700,000 tweets were extracted daily. Punctuation marks were not eliminated in the preprocessing step, and the entire contents were stored intact as tweets in general express user emotions.

[1] http://en.wikipedia.org/wiki/Main_Page
[2] http://wiki.dbpedia.org/Downloads2014
[3] https://about.twitter.com/what-is-twitter/

4 Implementation of Database Based on the Global Standard Using UIMA

Documents collected by the Scientific Technology Text Crawling System were stored in the CAS format of UIMA after the preprocessing procedure. UIMA is an open Apache source project that defines the common systematic structures of software that analyzes large volumes of unstructured information in order to discover knowledge that is relevant to an end user. CAS is the defined form of structures that express the feature and annotation used in UIMA. CAS is redefined and used for meeting the characteristics of the collected metadata information. As information, which is collected in the CAS format of UIMA, is expressed as the structure of the global standard, it can be used as the input data for the engine that extracts unstructured information using UIMA. The text documents collected and preprocessed in the CAS format through this study were transmitted to the Hadoop-based information extraction system [12].

Table 1. Status of archiving big data from text related to scientific technology

Type of scientific technology text	Number of documents
Web News	5,656,465
Article(Korean)	962,984
Article(English)	13,744,480
Patent	9,427,117
Wikipedia	4,004,000
Tweet	204,445,587

Table 1 shows the current status of the database data that have been collected and implemented thus far. 5,656,465 documents were collected from news websites, including articles published between 2001 and 2014. Patent data published between 2001 and 2013 were collected in bulk. About 200 million tweets posted since January 2014 were collected. Articles from news websites, foreign papers, and tweets were collected real-time.

5 Conclusion

This study analyzed a practical method of collecting multiple types of big data texts related to scientific technology and of implementing a database. The system collected various types of information, such as patents, data from news websites, Wikipedia content, and social media content, and systematically implemented a text database and distributed it in the CAS format of UIMA, the global standard format.

In the future, we intend to study ways to apply the characteristics of the real-time data collection system, which is currently being applied only to the crawlers of Web news, foreign papers, and social media, to other types of information. Text information will also be used for analyzing and predicting the trends of scientific technology, which is necessary to help researchers to improve their research.

Acknowledgments

This work utilized scientific and technical contents constructed as part of the "Establishment of the Sharing System for Electronic Information with Core Science and Technology" project (K-14-L02-C01-S02) and the "S&T Contents Construction and Service" project (K-14-L02-C01-S04). We would like to thank the Department of Overseas Information and Department of NDSL Service for providing certain contents used in this study.

References

1. Big Data, Big Impact: New Possibilities for International Development, `http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf`
2. IBM, Innovate with new analytics tools and technologies, `http://www.ibm.com/analytics/hk/en/what-is-smarter-analytics/innovate-with-analytics-tools.html`
3. Chen, H., Chiang, R., Storey, V.: Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, vol. 36, no. 4, pp.1165–1188 (2012)
4. Hwang, M., Cho, M., Hwang, M., Lee, M., Jeong, D.: Technical Terms Trends Analysis Method for Technology Opportunity Discovery. INFORMATION-AN INTERNATIONAL INTERDISCIPLINARY JOURNAL, vol. 17, no. 3, pp. 877–883 (2014)
5. Ferguson, M.: Architecting A Big Data Platform for Analytics. A Whitepaper Prepared for IBM (2012)
6. Burner, M.: Crawling towards eternity: Building an archive of the World Wide Web. `http://www.webtechniques.com/archives/1997/05/burner/` (1997)
7. Soumen, C., Berg, M., Dom, B.: Focused crawling: a new approach to topic-specific Web resource discovery. Computer Networks, vol. 31, pp.1623–1640 (1999)
8. Aggarwal, C.C., Al-Garawi, F., Yu, P.S.: Intelligent crawling on the World Wide Web with arbitrary predicates. Proceedings of the 10th international conference on World Wide Web, pp.96–105 (2001)
9. Hafri, Y., Djeraba, C.: High performance crawling system. Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval, pp. 299–306 (2004)
10. Exposto, J., Macedo, J., Pina, A., Alves, A., Rufino, J.: Geographical partition for distributed web crawling. Proceedings of the 2005 workshop on Geographic information retrieval, pp. 55–60 (2005)
11. Apache UIMA, `https://uima.apache.org/`
12. Um, J., Jeong, C., Choi, S., Lee, S., Jung, H.: Fast Big Textual Data Parsing in Distributed and Parallel Computing Environment. Mobile, Ubiquitous, and Intelligent Computing, pp.267–271 (2014)