Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2013, Article ID 525687, 6 pages
http://dx.doi.org/10.1155/2013/525687

*Research Article*

# Distributed and Parallel Big Textual Data Parsing for Social Sensor Network

**Jung-Ho Um,[1] Chang-Hoo Jeong,[1] Sung-Pil Choi,[1] Seungwoo Lee,[1] Hwan-Min Kim,[2] and Hanmin Jung[1]**

[1] *Department of Computer Intelligence Research, Korea Institute of Science and Technology Information, 245 Daehakno, Yuseong-gu, Daejeon 305-806, Republic of Korea*
[2] *Department of Overseas Information, Korea Institute of Science and Technology Information, 245 Daehakno, Yuseong-gu, Daejeon 305-806, Republic of Korea*

Correspondence should be addressed to Sung-Pil Choi; spchoi@kisti.re.kr

Received 30 August 2013; Revised 19 November 2013; Accepted 20 November 2013

Academic Editor: Hwa-Young Jeong



Recently, due to the popularization of the smartphone and social network service (SNS), many SNS users write their opinions for social events. According to these social phenomena, social sensor network which analyzes social events by utilizing those users' text data is proposed. Parsing is essential module to analyze user's text contents because it gives the understanding of semantics by extracting the words and their classes from texts. However, parsing requires much time because it needs to analyze all context information from the users' text. In addition, as users' text data are generated and transferred in streaming, the required parsing time increases too. This situation occurs that it is hard to parse the text on the single machine. Therefore, to drastically enhance the parsing speed, we propose distributed and parallel parsing system on the MapReduce. It applies the legacy parser to the MapReduce through loose coupling. Also, to reduce communication overheads, the statistical model used by the parser is resided on local cache in each mapper. The experimental result shows that the speed of proposed system is 2–19 times better than that of the legacy parser. As a result, we prove that the proposed system is useful for parsing text data in social sensor network.

## 1. Introduction

Currently, as a result of the development of smartphone device techniques and active use of the social network service (SNS), many people use SNS smartphone applications. In the United States, there are 30 million smartphone users in 2011 and had increased by million users per week in the last quarter of 2012 (http://tech.fortune.cnn.com/2012/03/07/u-s-smartphones-inching-toward-1-million-per-week/). In addition, Twitter, a representative SNS provider, has about 50 million cases of uploading per day and FaceBook, another provider, has about 60 million cases of uploading (http://allfacebook.com/twitter-facebook-status_b11613). Smartphone users post writings about their daily life or share major social events or issues using SNS such as Twitter, Facebook, U-tube, and flicker.

Many studies have been made to find out social issues or to solve scientific problems through these smartphone users' big text data [1–11]. For example, to detect radiation values generated by the japan's Fukushima Daiichi nuclear disaster, crowd sourced real-time radiation maps are developed by providing detecting radiation information from the smartphone users. Another example is constructing the citizen sensor network for the scientific discovery by offering volunteers' location information with their pictures. Those kinds of system are called social sensor network or specifically citizen-sensor network [4–11]. The analysis and monitoring system based on this social sensor network requires parsing to analyze the users' postings. Therefore, a system that can parse several thousand million of sentences within a short time is essential for the social sensor network to analyze social phenomena.

However, parsing requires a lot of time to run for a large number of documents because it considers semantics for wide context range of the sentence to enhance precision and recall value. This type of processing proportionally

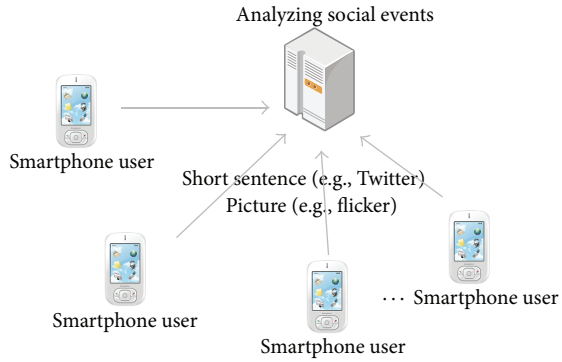


Figure 1: The system architecture of social sensor network.

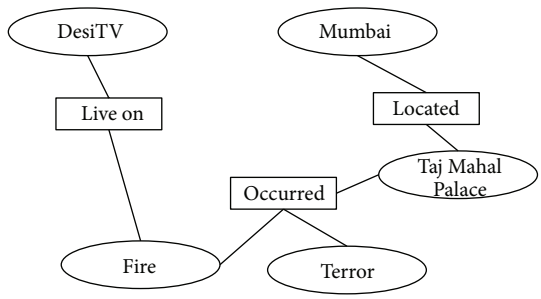


Figure 2: Sentence analysis.

affects the execution time with increasing in the number of documents. Therefore, a parsing system running on a distributed and parallel environment needs to be developed to parse massive numbers of documents concurrently. In this paper, we propose a parsing system that applies a Stanford parser to the MapReduce framework in order to extract information very fast. The reason why we use the Stanford parser is that it is one of the-state-of-the-art parsers and has high precision and recall in parsing documents.

The paper is organized as follows. Section 2 introduces related work. We report on the design of the proposed distributed and parallel text parsing system in Section 3. Section 4 describes the experimental results. Finally, in Section 5, we give the conclusion.

## 2. Background

To understand relationship between parsing and social sensor network, this section presents the social sensor network. Social sensor network means that the infra for analysis and collection of data from various user upload contents to detect social issues or to solve scientific problems. In the social sensor network, all smartphone users act as separate sensors by uploading their writings and photographs to the centralized server. This server finds current focusing social issues by analyzing user's context (see Figure 1). Social sensor network can analyze and monitor social phenomena more exactly since social sensor network reflects many people's opinion directly. Therefore, social sensor network is very useful to build intelligent social analyzing services [12–15].

For this, a technology called natural language processing that can analyze people's writings is needed. As shown in the example presented in [4], let us assume that a user posts a statement "mumbai taj occurred fire live on desitv" on Twitter. When this sentence is parsed, such words as "Mumbai," "Taj," "DesiTV," "live on," "fire," and "occurred" would be extracted. With dictionaries that specify places and institutions, it would be identified that "Mumbai" is a name of a place; "Taj" refers to "Taj Mahal Palace;" and the event is being broadcast by an Indian TV channel called DesiTV. If another user posts a statement on the fire saying "The fire in the Taj Mahal was turned out to be the terror," such words as "Taj Mahal," "fire," and "terror" are extracted for analysis and then linked to the analysis data for the previous post as shown in Figure 2. As shown in the figure, it can be identified that the fire in the Taj Mahal Palace in India was an act of terrorism. In this way, it is possible to understand the nature of social issues and events accurately through the writings posted by people. In addition, through such analysis, it is possible to determine how many postings people upload about similar issues and how much interest these issues attract.

## 3. Related Works

To parse user upload data from social sensor network, it needs two requirements. One is the high precision and the other is rapid speed. Therefore, it describes the Stanford parser related to the high precision and presents MapReduce framework related to the rapid speed. In this section, we introduce the Stanford parser [16] and the MapReduce framework [17]. These are used by the proposed system in distributed and parallel environments. First, the Stanford parser, proposed by the NLP lab of Stanford University in the 1990s, enhances precision and optimizes the performance by using the probabilistic context-free grammars (PCFG) model [16]. A PCFG is a context-free grammar in which each formal grammar is augmented with a probability. The Stanford parser is currently released as an open source program. The Stanford parser analyzes sentences using the PCFG model and notes the subjects, objects, and related verbs as a form of dependency tree structure. The PCFG model is also used for inference from sensing data's relationships in wireless sensor network [18]. However, many studies that have considered the Stanford parser focus on enhancing the performance of the parsing algorithm. The Stanford parser runs on the single machine so that it has much time to parse huge users' text from social sensor network. Hence, it is necessary to reduce the parsing time to quickly detect social event or issues in social sensor network.

On the other hand, the MapReduce framework is a parallel programming model proposed by Dean and Ghemawat in 2003 [17]. It consists of a user-defined Map function and a Reduce function. These two functions reside on each server, in order to allow the servers to process data in parallel (see Figure 3).

As can be seen in Figure 3, the input data is equally split ant then assigned to each server. The Map function processes data locally on each of the servers and the Reduce function

```
(1)    Class Mapper
(2)      Method Setup()
(3)        DistributedCache.add(PCFGmodel)
(4)        StanfordParser.setModel(PCFGmodel)
(5)      Method Map(document d)
(6)        sentence <- SetenceSeparator(d)
(7)        dependencyTree <- Parser.parsePCFG(sentence)
(8)        emit(d.id, dependencyTree)
(9)    Class Reducer
(10)     Method Reduce(d.id, Iterable<dependencyTree>)
(11)         for each dependecyTree dt
(12)              output+=dt
(13)          emit(d.id, output)
```

Algorithm 1

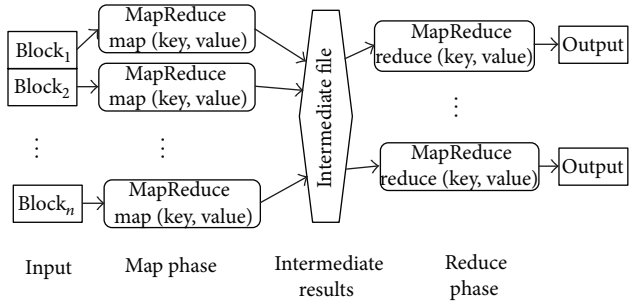


Figure 3: The MapReduce framework.

merges the computed data following a user-defined process. MapReduce can analyze big data efficiently. Therefore, it has been applicable to process social sensor network data. In addition, MapReduce supports fault tolerance so users only focus on developing Map and Reduce functions to analyze data in parallel.

## 4. Proposed Distributed and Parallel Parsing System

To design distributed and parallel parsing system for social sensor network data, we firstly decide on the way how to process the data in distributed and parallel environment. It has two ways which are tightly coupled and loosely coupled integration between distributed and parallel system and parser. At the first, tightly coupled integration can be developed by processing the partial part of the sentence in parallel. However, the job's unit is very tiny so that it is not adapted for data intensive computing. The reason is that the additional communication overhead can occurr since the data size for the one job is generally a number of megabytes. On the other hand, loosely coupled integration can be processed by sentences so that we can archive scalability in terms of number of servers. In addition, it has an advantage that it skips the step of splitting sentence.

In order to extract information from massive numbers of documents from SNS users' contents in citizen-sensor network, we have designed a distributed and parallel textual parser which is implemented by applying the parser to the MapReduce framework. The proposed system architecture is shown in Figure 4.

The system stores the input and output data, such as documents and parsed sentences, in the hadoop file system [19]. The model that is used by the parser using machine learning technique is also loaded into the hadoop file system because the PCFG model is necessary for all mappers, allowing them to parse sentences. Therefore, the model is loaded as a distributed cache for sharing all of the mappers. The work flows of the mapper and the reducer are described as follows. First, the mapper separates documents into sentences and then parses the sentences by calling the PCFG parsing module of the Stanford parser. At that time, the Stanford parser needs the PCFG model. For this reason, the PCFG model is stored in the distributed cache. After the completion of parsing, the map function writes the parsed sentence information in a dependency tree form. This format can be used to recognize relations between words or phrases. Next the reducer merges the dependency tree structure for each sentence and then writes final results.

The pseudocodes of proposed system are as Algorithm 1. First, Mapper class consists of Setup and Map functions. In Setup function, the PCFG model is loaded to the distributed cache of the each mapper (Lines 3-4). Map function separates sentences from the input documents (Line 6). After that, Stanford parser library is called to parse for each sentence and writes to intermediate file (Line 7-8). Next, Reducer class has Reduce function. This merges the results ordered by document id and writes output data (Line 10–13).

Proposed system has three advantages. First, it reduces the parsing time because the system analyzes massive number of documents concurrently on distributed and parallel environments, while legacy parsers require a lot of parsing time because they analyze documents sequentially. Second, the proposed system can maintain the same high precision performance as the Stanford parser applied to the proposed system. Finally, the system has high portability. The reason for this is that if users want to change the parser, the system can easily be changed by modifying only the parser calling

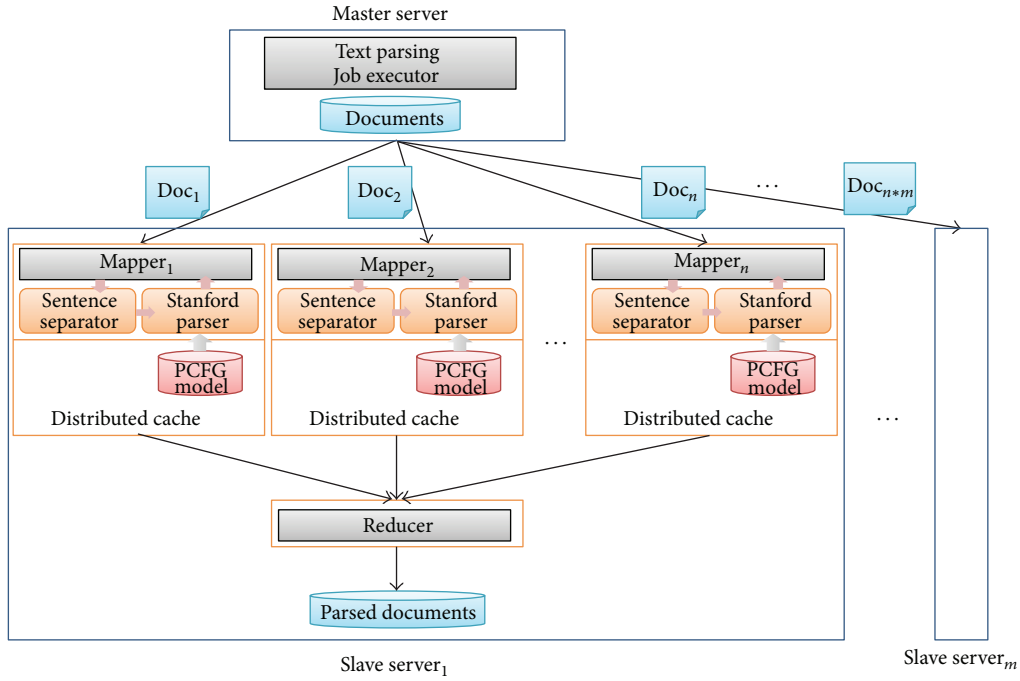


FIGURE 4: The system architecture.

API part with replacing the Stanford parser with the legacy parser.

## 5. Experimental Results

To evaluate the performance of the proposed system for parsing user's sentences from citizen-sensor network, we consider two different HW environments such as single server and nineteen servers. On the single server, we compare the running time of the legacy Stanford parser with that of the distributed Stanford parser applied to the MapReduce framework. We also evaluate the running time of the distributed Stanford parser on nineteen servers. Servers consist of eight cores of Intel i7, 32 GByte memories and 2TB storages; we use hadoop 0.20.203 and the Stanford Parser 2.0.4. The first data set consists of 10,000 paper abstracts from NDSL owned by KISTI NDSL.

The abstract of NDSL is generated by parsing XML data which stores metadata of NDSL. Among them, we randomly choose. The reason why we use abstracts from NDSL is that the lengths of the sentences are quite similar to the length of SNS users' writings. The second data set is for evaluating the scalability of distributed and parallel parsing system. It has 16,000,000 sentences from collecting technical web articles, papers, and patents at 2012 by KISTI. We evaluate the running time from the step of the sentence separation to the step of parsing sentences using the distributed Stanford parser. The experimental results are shown in Figure 5. The java heap memory size for the experiments is set to 4 GB. For the single server, the execution time of the proposed system is almost half that of the legacy Stanford parser, when the numbers of mappers and reducers are four and one, respectively. Even though the system uses four mappers, the performance gain is only double. The reason for this is that the system requires starting time to initialize the MapReduce framework and additional time to merge the results. In this experiment, we found that the speed is proportional to half the number of mappers. The results for the distributed Stanford parser, where that parser is running on 19 servers, follow our predictions. This means that the proposed system has an advantage in terms of scalability.

We evaluate the parsing time by increasing sentences from 1,600,000 to 16,000,000. As shown in Figure 6, the parsing time increases proportionally to the number of sentences. It means that the proposed system has scalability in terms of data size.

## 6. Conclusion

In this paper, a parser system that can be used for analyzing users' sentences in the social sensor network was proposed. To this end, the existing Stanford parser with high performance in terms of accuracy was applied to MapReduce framework for distributed and parallel processing. The PCFG model used by the Stanford parser was loaded to each mapper, using distributed cache, to maximize the locality of calculation. In addition, in order to increase the system's portability, the proposed parser was implemented to call and use the Stanford parser library through loose coupling, rather than linking the Stanford parser and MapReduce through tight coupling. In the performance evaluation, the proposed system showed the performance two times that of the existing parser, in a single machine, through the parallelization of MapReduce framework, parallelization of

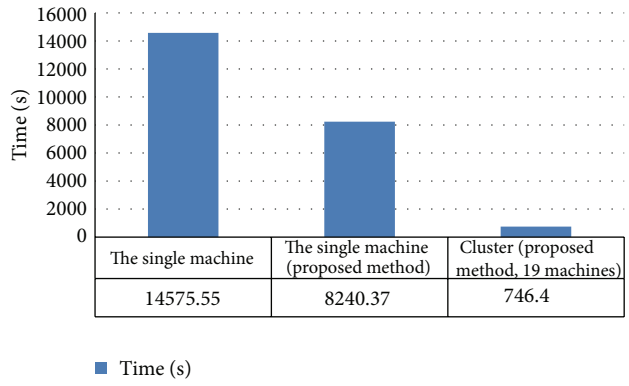


Figure 5: Experimental results.

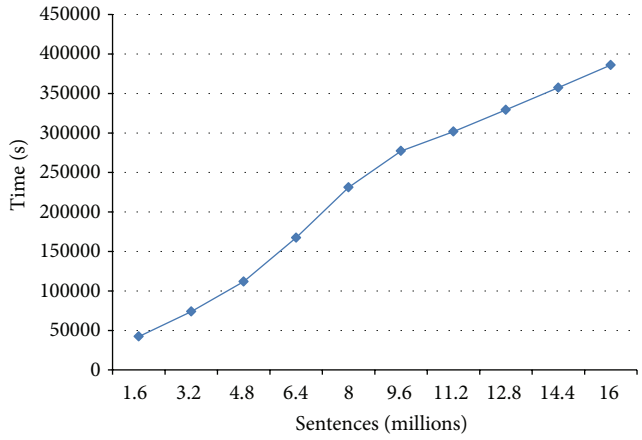


Figure 6: Experimental results.

the system. Furthermore, the parser showed 19 times the performance of a single machine, which covered 19 servers. As a result, we found that the proposed system is useful for parsing text data in social sensor network.

For future studies, performance evaluation for the proposed parsing technique will be conducted by analyzing actual Twitter SNS data observed by the social sensor network, and technically specialized sentences as well as ordinary user statements will also be analyzed based on various datasets [20, 21] such as web documents, patents, and papers. With regards to the distributed parallel processing, a study for finding out the optimized data processing environment while varying the MapReduce execution environment will be conducted.

## Acknowledgments

This work utilized scientific and technical contents constructed through "Establishment of the Sharing System for Electronic Information with Core Science and Technology" Project (K-13-L02-C01-S02). And authors declare that there is no conflict of interests regarding the publication of this paper.

## References

[1] D. de Roure, N. R. Jennings, and N. R. Shadbolt, "The semantic grid: past, present, and future," *Proceedings of the IEEE*, vol. 93, no. 3, pp. 669–680, 2005.

[2] P. Bellini, I. Bruno, D. Cenni, and P. Nesi, "Micro grids for scalable media computing and intelligence in distributed scenarios," *IEEE MultimediaIssue*, vol. 19, no. 2, 2012.

[3] M. Cannataro, P. H. Guzzi, and A. Sarica, "Data mining and life sciences applications on the grid," *Data Mining and Knowledge Discovery*, vol. 3, no. 3, pp. 215–238, 2013.

[4] A. Sheth, "Citizen sensing, social signals, and enriching human experience," *IEEE Internet Computing*, vol. 13, no. 4, pp. 87–92, 2009.

[5] M. N. Kamel Boulos, B. Resch, D. N. Crowley et al., "Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples," *International Journal of Health Geographics*, vol. 10, p. 67, 2011.

[6] M. Nagarajan, K. Gomadam, A. P. Sheth, A. Ranabahu, R. Mutharaju, and A. Jadhav, "Spatio-temporal-thematic analysis of citizen sensor data: challenges and experiences," in *Web Information Systems Engineering-WISE*, pp. 539–553, 2009.

[7] D. Zhang, B. Guo, and Z. Yu, "The emergence of social and community intelligence," *Computer*, vol. 44, no. 7, Article ID 5719570, pp. 21–28, 2011.

[8] M. Nagarajan, A. Sheth, and S. Velmurugan, "Citizen sensor data mining, social media analytics and development centric web applications," in *Proceedings of the 20th International Conference Companion on World Wide Web (WWW '11)*, pp. 289–290, April 2011.

[9] D. Zhang, B. Guo, B. Li, and Z. Yu, "Extracting social and community intelligence from digital footprints: an emerging research area," *Ubiquitous Intelligence and Computing*, vol. 6406, pp. 4–18, 2010.

[10] D. Villatoro and J. Nin, *Citizens Sensor Networks*, vol. 7685 of *Lecture Notes in Computer Science*, 2013.

[11] J. K.-Y. Ng, "Ubiquitous healthcare: healthcare systems and applications enabled by mobile and wireless technologies," *Journal of Convergence*, vol. 3, no. 2, 2012.

[12] J. Kim, S. Lee, D. H. Jeong, and H. Jung, "Semantic data model and service for supporting intelligent legislation establishment," in *Proceedings of the 2nd Joint International Semantic Technology Conference*, 2012.

[13] V. Viswanathan and I. Krishnamurthi, "Finding relevant semantic association paths through user-specific intermediate entities," *Human-Centric Computing and Information Sciences*, vol. 2, p. 9, 2012.

[14] B. J. Oommen, A. Yazidi, and O. C. Granmo, "An adaptive approach to learning the preferences of users in a social network using weak estimators," *Journal of Information Processing Systems*, vol. 8, no. 2, pp. 191–212, 2012.

[15] J. Ortiz, A. García-Olaya, and D. Borrajo, "Using activity recognition for building planning action models," *International Journal of Distributed Sensor Networks*, vol. 2013, Article ID 942347, 10 pages, 2013.

[16] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423–430, 2003.

[17] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[18] S. C. Geyik and B. K. Szymanski, "Event recognition in sensor networks by means of grammatical inference," in *Proceedings of the 28th Conference on Computer Communications (IEEE INFOCOM '09)*, pp. 900–908, April 2009.

[19] HDFS (hadoop distributed file system) architecture, 2009, https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html.

[20] S. Shin, J. Um, S.-K. Song, S.-P. Choi, and H. Jung, "uLAMP: unified linguistic assets management system," in *Proceedings of the 2nd Joint International Semantic Technology Conference*, 2012.

[21] D. Seo, M. N. Hwang, S. Shin, and S. Choi, "Development of crawler system gathering web document on science and technology," in *Proceedings of the 2nd Joint International Semantic Technology Conference*, 2012.