Association for Information Systems
AIS Electronic Library (AISeL)

WHICEB 2015 Proceedings | Wuhan International Conference on e-Business

Summer 6-19-2015

OSBIA: Open Source Business Intelligence Analytics System Based on Domestic Platform

Shiwei Zhao
Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, China, xiaoyaow@126.com

Zewen Cao
Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, China

Wensen Liu
Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, China

Follow this and additional works at: http://aisel.aisnet.org/whiceb2015

Recommended Citation
Zhao, Shiwei; Cao, Zewen; and Liu, Wensen, "OSBIA: Open Source Business Intelligence Analytics System Based on Domestic Platform" (2015). *WHICEB 2015 Proceedings*. 25.
http://aisel.aisnet.org/whiceb2015/25

This material is brought to you by the Wuhan International Conference on e-Business at AIS Electronic Library (AISeL). It has been accepted for inclusion in WHICEB 2015 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

OSBIA: Open Source Business Intelligence Analytics System Based on Domestic Platform

Shiwei Zhao[1*], *Zewen Cao*[1], *Wensen Liu*[1]
[1]Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, China

Abstract: Nowadays, online comments and other textual data become more and more significant for business intelligence service. However, there is blank in the area of IS based on domestic platform at present. We designed and implemented OSBIA: an open source business intelligence analytics system based on domestic platform. OSBIA system concentrates on analyzing open source textual intelligence for the business purpose and adopts self-designed distributed crawler system so that a closed circle is formed from intelligence collection to analysis process and push service. For the efficiency of OSBIA, the improved parallel OPTICS algorithm (IPOPTICS) is described in this paper, which effectively solves the problem that OPTICS is limited by its parameter- radius of neighborhood. And we also illustrate one typical application for market research: "consumer's comments discovery" based on Stanford parser and IPOPTICS algorithm and the like. The results of experiences show that the OSBIA system is much faster than the original OPTICS and suitable for large scale textual intelligence analysis.

Keywords: intelligence analysis system, domestic platform, text mining, cloud computing

1. INTRODUCTION

The era of big data brings both opportunities and challenges for intelligence analysis [1]. Western countries have already paid large attention to intelligence service based on open-source data and developed many technologies about collecting, processing and mining the open-source intelligence, in order to promote their intelligence analysis ability. America, especially, proposed many strategies about the big data. DRAPA proposed its big data research project: XDATA to improve its big data processing capacity for the intelligence service [2-5].

Since 1999, Chinese government has set up several strategies to support domestic basic software based on open-source software. As developing basic software such as operating system is listed as an important event in " National Principle of Scientific and Technical Development (2006 -2020）"[6], domestic platform operating system has gradually become the prior choice for many corporations and departments. It would be safer, if the intelligence analysis system is based on domestic platform operating system.

However, most intelligence analysis systems are based on Windows. For instance, there are many media monitor applications, such as Europe Media Monitor[7], NRTIM(Near Real Time Information Mining in Multilingual News)[8], STALKER[9], and many business intelligence software, like Actuate, JasperSoft, OpenI, Palo, Pentaho and SpagoBI[10]. It is really urgent that an intelligence analysis system based on domestic platform operating system is required.

OSBIA (open source business intelligence analytics) system based on domestic platform is designed and implemented. OSBIA is aimed to provide efficient, accurate and convenient services to analyze massive open-source business textual data from the Internet. It offers strong text analysis services, supports complex requests for transaction analysis. And more importantly, for the sake of the security and utility, it is compatible with the NeoKylin operating system and Chinese domestic databases, and adapts the J2EE framework.

This paper is structured as follows: related work is shown in Section 2; Section 3 describes the framework

* Corresponding author. Email: xiaoyaow@126.com(Shiwei Zhao)

of OSBIA by illustrating the function of different parts; IPOPTICS is introduced in Section 4; Section 5 presents the typical application of OSBIA and the performance followed by Section 6 giving the conclusions.

2. RELATED WORK

Nowadays, the open source intelligence, such as news and consumers` opinion of the product, becomes more and more useful and valuable. Liu *et al* [11, 12] presents TIARA, which combines text mining and interactive visualization to help users explore and analyze large collections of text. Marlies *et al* [13] designed Facebook watchdog for detecting online grooming and bullying activities. V.K. singh *et al* [14] analyzed thematic blog data through combine of sophisticated IR and NLP for finding out important inferences about discrimination, abuse and sexual crime against women. When it comes to business service, Alan *et al* [15] found out the vehicle defect through analyzing the conversation on online discussion forums; Lipika *et al* [16] proposed a framework for text-driven analysis of multi-structured data for enterprise intelligence.

Clustering algorithm is the basis of big data processing and offer general analysis for textual data. It is significant in the later processing, such as topic detection, automatic summarization and retrieval acceleration [17-19].

3. FRAMWORK

OSBIA mainly consists of 4 parts: Data Acquisition uses distributed crawler to gain massive data; Data Storage stores the raw data and some pretreatment data; Computational Analysis provides several distributed analysis algorithm for forward treatment; Human Interface offers applications and visualized intelligence to users. OSBIA which is mainly based on NeoKylin-the Chinese domestic platform operating system operating system, uses Message Middleware to transmit information and utilizes Security Middleware to ensure the system safety. Eucalyptus is used to set up cloud environment. The framework is showed as follows:

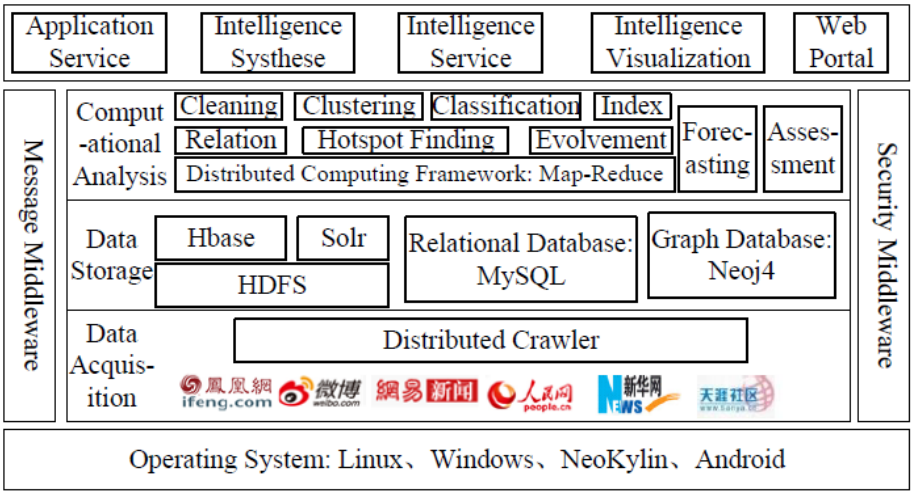

Figure 1. System architecture

3.1 Data collection

The distributed crawler which was self-designed is deployed on the cloud platform. After getting the request from users, it obtains massive text information through the Internet. For instance, users` opinion would be gotten from Weibo, Tianya and other Chinese social network sites, and news would be gained from the NNS, ifeng.com and other news websites. The crawled data would be stored in the distribute database: Hbase.

3.2 Data storage

OSBIA utilizes HDFS and Hbase to store and control the raw crawled data, and uses Solr to create index and retrieval for the data. Some processed data would be stored in the relation database, for example, Oracle, MySQL, Dameng, Kingbase and some generated diagram would be stored in Neoj4.

3.3 Computational analysis

Computational analysis is the main component of OSBIA, as it offers all data mining algorithm for the subsequent analysis. After the cleaning step, the textual data would be clustered or classified for the different purposes, and then indexed by Solr. Then there could be several choices, such as relation analysis: finding the implicit relation between different data, hotspot finding: figuring out the hotspot in a group of data, and evolvement analysis: understanding the development of specific topic. All these algorithms are used in the distributed computing framework-MapReduce. In addition, some forecasting and assessment algorithm are also provided for future analysis.

3.4 Human interface

This part gives users some services which they could get required information by different type, like diagram, table or text. Application service provides users with various customized application, which would be present in Section 5. Intelligence synthesis offers users a comprehensive result of a various range of data, which makes users have a more clear and precise inspect in these data. Intelligence Service gives users diverse choice of different text mining algorithm and combination of them, in order to satisfy their needs. Users could get some visual result from intelligence visualization, such as generated relation diagrams and other graphs which illustrated the properties of specific intelligence. All the service could be gotten from a web portal.

4. KEY TECHNOLOGY

OPTICS is one kind of density clustering algorithm. The core concept is to use number of neighbor points in neighborhood size (ε) of specific point to calculate the density of this point [20]. In practice, the evaluation of ε influences the effectiveness of the algorithm significantly. Most proposed improved algorithm, however, did not optimize the evaluation of ε [21, 22]. Johannes [23] proposed fast OPTICS algorithm based on random projections, which replaced ε with the average distance after projections. But for textual data, the random projections are complex and time-consuming because of its high dimension. As a result we proposed an improved parallel OPTICS: IPOPTICS based on the algorithm proposed by Johannes.

Firstly, the whole textual data set is divided into several smaller sets by partition step. Then improved OPTICS is utilized in every single set, which promote the efficiency of the algorithm as the distance between data from different groups would not be calculated

Partition is the first step of the algorithm. In the original algorithm, the points are projected onto a random line, and a point that has been projected on the line is chosen uniformly at random which is unsuitable for textual data, as the calculation is complex in the high-dimensional space. As a result, we proposed another partition algorithm, which used the distance between two points to split the points. To be more specific, one point is chose randomly and the distances between this point and the other points are calculated in a parallel form. Then, sort the distance to find out the largest D-value between adjacent distances. We split it recursively into two parts until the size of the pint set is at most *minSize*, where *minSize* is a parameter of the algorithm. Use the result to split the points into two disjoint sets. The algorithm is shown as follows:

Algorithm: Partition
Input:
minSize: the threshold value of the algorithm
S: the entire textual data set
Output:
G: the set of data sets after splitting
Method:
(1) *G*:={S};
(2) **Repeat**
(3) Choose a set Q randomly whose size is more than *minSize* and choose a text a from Q
(4) Calculate the distance between a and all others and sort the result
(5) Find out the largest D-value of the queue, and the corresponding points b, c
(6) Split points whose distance are less than that of b and the point b into one part Q1, the others into another part Q2
(7) Remove Q from *G* and put Q1, Q2 into *G*
(8) **Until** the size of every set in *G* is no more than *minSize*

We use the average distance $D_{avg}(a)$ between point a and its neighbor points for a parameter *dPts* to replace ε. The algorithm is exhibited as follows:

Algorithm: Calculation
Input:
dPts: the maximum number of neighbor points
Q: the textual data set
Output:
$D_{avg}(a)$: the average distance of point a
Method:
(1) Choose a point a randomly from Q
(2) Calculate the distance between point a and other points, and sort the points by the distance in ascending order
(3) Choose the top *dPts* the calculated distances to form list D
(4) **If** the D-value between two sequential values of D (b, c and b is smaller than c) is much larger than other D-value between two sequential value of D
(5) Chang *dPts* to the serial number of b
(6) Calculate the sum of top *dPts* values of D: dis
(7) $D_{avg}(a)$= dis/*dPts*

The IPOPTICS algorithm is exhibited as follows:

Algorithm: IPOPTICS
Input:
minSize: the threshold value of the algorithm
S: the entire textual data set
dPts: the maximum number of neighbor points
Output:
Result: the sequence of all points and a distance for each point.
Method:
(1) Use Partition algorithm to divide S into several data sets

(2) In every single data set, use the Calculation algorithm to get the average distance $D_{avg}(a)$ of point a and replace ε with $D_{avg}(a)$. The definition of the reachability distance for a point a and a reachable point b from a is the same as for OPTICS

(3) Utilize OPTICS to produce the Result

IPOPTICS is easy for Map-Reduce. At partition step, Map function is used to divide the data set and Reduce function is used to sort the results. And at clustering stage, Map function is used to calculate the distance and segment the points and Reduce function is used to output the results. Besides, at every step, the datasets are distributed to several nodes for the forward processing.

5. TYPICAL APPLICATION & PERFORMANCE

OSBIA provides users with general text mining functions, such as text classification, text clustering, automatic abstraction, hot-topic discovery, theme extraction and relevance analysis. Moreover, it has developed several customized applications for special purpose, for instance, personnel relation exhibition, personnel behavior analysis, key individual identified. An application about supporting market research is presented in this section- "consumers' comments discovery".

Consumers' comments would be found out through analyzing the open source intelligence, such as news report and user views in the social websites and forums. Users' views of the new product or the service of one corporation would be clearly exhibited in different groups, which would help managers to draw up new marketing strategy.

In order to find out consumer's opinion, after pre-processing the text, Stanford parser was used to extract the key point of single comment and the results were stored in the database. Then, improved parallel OPTICS algorithm was utilized to segment comments into several clusters by the stored results. The entire data is collected in an xml format and a summary of it collected from the different sites from January 2013 to June 2014 is given in Table 1.

Table 1. Dataset

Topic	No. of blog posts	No. of news report	No. of comments in social websites	No. of comments in forums	Total size
Smart phone	1535	979	11234	12123	1127M
Computer	1224	1007	10342	11278	922M
Cloth	1634	1242	13242	12134	1229M
Cosmetic	1123	1078	10253	10098	819M

The F1 score of the improved algorithm was 87.29% (the recall rate was 86.8% and the precise rate was 87.8%). The experience was performed on the 3 nodes Hadoop clusters, and the name node consisted of 8 Intel Xeon E52603 CPU, 32GB main memory and 2.51T hard drive and the data node consisted of 4 Intel Xeon I34130 CPU, 4GB main memory and 500GB hard drive. Fig.5 shows that the system had good performance and the algorithm had good extension. We also utilized the traditional OPTICS algorithm for the consumers` comments discovery. The results showed that it cost more than 1hour to process 500M data, which means that the IPOPTICS is much efficient than the OPTICS.

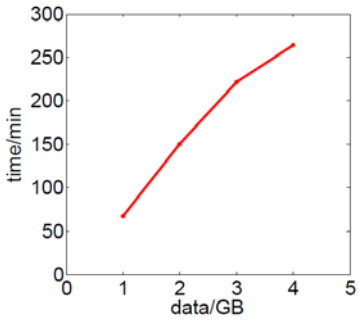

Figure 4. The efficiency of consumers' comments discovery

6. CONCLUSIONS

All involved middle wares and databases are compatible with the domestic platform operating system. As for operating platform, servers and clients use NeoKylin operating system. In terms of database, OSBIA is compatible with Oracle, SQL Server and Chinese domestic databases, such as DaMeng, Kingbase and KSTORE. When it comes to could platform, Eucalyptus guarantees the security because of the inside key certificate and outside firewall. In addition, OSBIA forms a closed circle from the open-source intelligence acquisition to analysis and offering customized intelligence analyzed result, and all data and results are stored in the domestic database, which ensure the safety.

OSBIA is concentrated on analyzing the open-source textual intelligence. It adopts the distributed crawler to ensure the efficiency of web crawler and IPOPTICS which improves the efficiency of traditional OPTICS algorithm. And other distributed text mining algorithms make sure that OSBIA could handle massive dataset. Apart from offered several general text mining functions, users can freely utilize the existing text mining module for their own purposes, which ensures its feasibility. Moreover, managers can get information by several forms, like diagram, table and so on. For different request from managers, OSBIA provides customized server offering more specific and professional results. As a result, OSBIA is a high-performance, rich-functional system. Further works includes: deeper study on the information extraction and improvement on the Stanford parser.

7. ACKNOWLEDGEMENT

The authors are grateful to the Big Data & Social Computing Engineering Centre, for his help in setting up the system and collecting data, and his helpful feedback as well.

REFERENCES

[1] Li Guangjian, YANG Lin. (2012).Intelligence Analysis and Intelligence Technology in View of Big Data. Lib & Info, 6: 1-8.

[2] Xie Liuxiang, YANG Pei, Luan Xidao, *et al*. (2007).Internet Information Collection and Processing Technology. Comp Eng, 33(23): 205-207.(in Chinese)

[3] NATO. (2011). Open Source Intelligence Reader. http://www.oss.net/dynamaster /file_archive/ NATO OSINT Reader FINAL 11OCT02.Pdf

[4] Whitehouse. (2011). Big Data Press Release. http://www.whitehouse.gov/sites/default/files/micros-.ites/ostp/ big_data_ press_release_final_2. pdf

[5] Whitehouse. (2012). Big Data is a Big Deal. http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal.pdf

[6] The Central Government of the PR China. (2006). National principle of scientific and technical development (2006 -2020). http://www.most.gov.cn/kjgh/.

[7] Best, C. Goot, E. Blackler. (2005).Europe Media Monitor. Technical Report EUR22173 EN, European Commission.

[8] M. Atkinson, E. Van der Goot. (2009).Near Real Time Information Mining in Multilingual News. In: Don Felipe de Borbón, Juan Quemada, Gonzalo León, eds. Proceedings of the 18th International Conference: World Wide Web. Austin: ACM. 1153-1154.

[9] F. Neri, M. Pettoni, Stalker. (2008).A Multilingual Text Mining Search Engine for Open Source Intelligence. In: Anna Ursyn , Ebad Banissi, eds. Proceedings of IV'08. New York: IEEE. 314-320.

[10] Joaquim Lapa, Jorge Bernardino, Ana Figueiredo. (2014).A comparative analysis of open source business intelligence platforms. In: S. Asghar, J. Darmont, eds. ACM International Conference Proceeding Series. Austin: ACM. 87-92.

[11] Wei Furu, Liu Shixia, Song Yangqiu et al. (2010).TIARA: A Visual Exploratory Text Analytic System. In: Bharat Rao, Balaji Krishnapuram, Andrew Tomkins Qiang Yang, eds. Proceedings of KDD'10. Austin: ACM. 153-163.

[12] Liu Shixia, Michelle X. Zhou, Pan Shimei et al. (2012).TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis. ACM Trans on Intel Sys & Tech, 3(2): 1-28

[13] Marlies Rybnicek, Rainer Poisel, Simon Tjoa. (2013).Facebook Watchdog: A Research Agenda For Detecting Online Grooming and Bullying Activities. In: Loi Lei Lai, Daniel S. Yeung, eds. Proceedings of ICSMC'13. New York: IEEE. 2854-3861.

[14] V.K. singh, P. Waila, R. sadat et al. (2013).Computational Analysis of Thematic Blog Data for Sociological Inference Mining. In: Haibo HE, Cesare Alippi, eds. Proceedings of ISACII'13. New York: IEEE. 293-298

[15] Alan S., JIAO Jian, G.Alan Wang *et al*. (2012).Vehicle defect discovery from social media. Dec Sup Sys, 54: 87-97.

[16] Lipika Dey, Ishan Verma. (2013).Text-driven Multi-structured Data Analytics for Enterprise Intelligence. In: William K. Cheung, Rajshekhar Sunderraman, eds. Proceedings of ICWIIAT'13. New York: IEEE. 213-220.

[17] Zeng H-J, He Q-C , Chen Z *et al*. (2004).Learning To Cluster Web Search Results. In: Kalervo Järvelin, Peter Bruza, Gareth Jones, eds. Proceedings of SIGIR'04. Austin: ACM. 210-217.

[18] Song Qinbao, Shen Junyi. (2002).A Web Document Clustering Based on Association Rule. Jour of Soft, 13(3): 417-423.(in Chinese)

[19] Li Baoli, Yu Shiwen. (2003).Research on Topic Detection and Tracking. Comp Eng & App, 17: 8-12.

[20] M. Ankerst, M. M. Breunig, H. peter Kriegel et al. (1999).Optics: Ordering points to identify the clustering structure. In: Proceedings of SIGMOD`99. Austin: ACM.

[21] Zeng Yiling, Xu Hongbo, BAI Shuo. (2008).OPTICS-Plus for Text Clustering. Jour of Chi Info Proc, 22(1): 51-60.

[22] Md. Mostofa, Ali Patwary, Diana Palsetia1 et al(2013).Scalable Parallel OPTICS Data Clustering Using Graph Algorithmic Techniques. In: Gropp, William, Matsuoka, Satoshi, eds. Proceedings of ICHPCNS'13. New York: IEEE. 42-57.

[23] Johannes Schneider, Michail Vlachos. (2013).Fast Parameterless Density-Based Clustering via Random Projections. In: Arun Iyengar, Qi He, Jian Pei, Rajeev Rastogi, eds. Proceedings of CIKM`13. Austin: ACM. 861-866.