6. CONCLUSIONS

Keyword searching over Big Data acts as a vital role and basis for multiple applications. The ever increasing data size, data complexity and data heterogeneity provides both challenge and opportunity for industries as well as academic disciplines. In this paper, we propose a software infrastructure supported by the distributed back-end hardware clusters to address the problem of keyword searching over Big Data. In the future, we plan to establish and integrate the framework collectively and finally provide end users on-demand service. Our project is still under progress. In this paper we particularly focus on the hardware implementation part. We will leave the rest for future research. Further work needs to be done to establish an efficient algorithm to perform keyword searching over large amount of data. Although the current study is based on a small sample of the dataset , the findings suggest that by tuning the number of machines and the configurations of each machine, we can decrease the loading time and searching time substantially. In the future, based on the hardware foundation we build an Entity-Unit keyword searching algorithm using the same experimental setup.

7. REFERENCES

[1] C. Lynch, "Big data: How do your data grow," *Nature*, vol. 455, no. 7209, pp. 28–29, 2008.

[2] P. Joshi, I. Pathan, and A. Khan, "Keyword Generation for Search Engine Advertising," *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 6, pp. 367–373, 2014.

[3] C. Bizer, P. Boncz, M. L. Brodie, and O. Erling, "The meaningful use of big data: Four perspectives – four challenges," *SIGMOD Rec.*, vol. 40, pp. 56–60, Jan. 2012.

[4] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[5] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, "Grouplens: Applying collaborative filtering to usenet news," *Commun. ACM*, vol. 40, pp. 77–87, Mar. 1997.

[6] J. B. Schafer, J. Konstan, and J. Riedl, "Recommender systems in e-commerce," in *Proceedings of the 1st ACM Conference on Electronic Commerce*, EC '99, (New York, NY, USA), pp. 158–166, ACM, 1999.

[7] G. Li, B. C. Ooi, J. Feng, J. Wang, and L. Zhou, "Ease: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 903–914, ACM, 2008.

[8] J. Shi, D. Wu, and N. Mamoulis, "Top-k relevant semantic place retrieval on spatial rdf data," in *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, (New York, NY, USA), pp. 1977–1990, ACM, 2016.

[9] L. Bo, L. Xianglong, and W. Li, *The Next-Generation Search Engine: Challenges and Key Technologies*, pp. 239–248. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[10] W. Dong, Z. Lei, and Z. Dongyan, "Top-k queries on rdf graphs," *Information Sciences*, vol. 316, pp. 201–217, 2015.

[11] Y. T. Yu, L Chang, "Scalable keyword search on large data streams," *IEEE 25th International Conference*, pp. 1199–1202, 2009.

[12] G. Piao, S. showkat Ara, and J. G. Breslin, "Computing the semantic similarity of resources in dbpedia for recommendation purposes," in *Joint International Semantic Technology Conference*, pp. 185–200, Springer, 2015.

[13] J. P. Leal, V. Rodrigues, and R. Queirós, "Computing semantic relatedness using dbpedia," in *OASIcs-OpenAccess Series in Informatics*, vol. 21, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[14] L. U. Quilitz, Bastian, *Querying Distributed RDF Data Sources with SPARQL*, pp. 524–538. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[15] Apache Software Foundation, "Apache hadoop version 2.6.5," 2016. `https://hadoop.apache.org`.

[16] Apache Software Foundation, "Aparche spark version 2.1.0," 2016. `http://spark.apache.org/`.

[17] Apache Software Foundation, "Aparche hbase version 1.3.0," 2017. `https://hbase.apache.org/`.

[18] Amazon, "Amazon dynamodb," 2012. `https://aws.amazon.com/dynamodb/`.

[19] Apache Software Foundation, "Aparche cassandra version 3.10," 2016. `http://cassandra.apache.org/`.

[20] Y. Mehta and S. Buch, "Semantic proximity with linked open data: A concept for social media analytics," in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, pp. 337–341, April 2016.

[21] Apache Software Foundation, "Aparche hadoop hdfs," 2016. `http://hortonworks.com/apache/hdfs/`.

[22] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web*, pp. 697–706, ACM, 2007.

[23] Wikimedia Foundation, "Wikipedia," 2017. `https://en.wikipedia.org/wiki/Wikipedia`.

[24] University of Princeton, "Wordnet version 2.1," 2015. `https://wordnet.princeton.edu/`.

[25] www.geonames.org, "Geonames," 2017. `www.geonames.org`.

[26] Max Planck Institute for Informatics, "Yago: A high-quality knowledge base," 2017. `https://gate.d5.mpi-inf.mpg.de/webyago3spotlx/SvgBrowser`.

[27] U. S. Administration, "Data.gov," 2017. https://www.data.gov/.

[28] A. Passant, "Measuring semantic distance on linking data and using it for resources recommendations.," in *AAAI spring symposium: linked data meets artificial intelligence*, vol. 77, p. 123, 2010.