IJITKM 0973-4414

International Journal of IT & Knowledge Management (ISSN: 0973-4414) www.csjournals.com
Volume 11 • Issue 2 pp. 81-86 Jan 2018 - June 2018 Published on 12-04-18

Knowledge-Based & Corpus-Based Methods for Evaluation of Semantic Relatedness of Concepts in Knowledge Graphs

Mohd Yousuf Ali[1] and Md Ateeq Ur Rahman[2]
[1]Research Scholar, Dept. of Computer Science & Engineering, SCET, Hyderabad
[2]Professor and Head, Dept. of Computer Science & Engineering, SCET, Hyderabad

Abstract: This paper presents a way for measuring the linguistics similarity between ideas in data Graphs (KGs) like WordNet and DBpedia. Previous work on linguistics similarity ways have centered on either the structure of the linguistics network between ideas (e.g. path length and depth), or solely on the data Content (IC) of ideas. we have a tendency to propose a linguistics similarity technique, specifically wpath, to mix these 2 approaches, mistreatment IC to weight the shortest path length between ideas. typical corpus-based IC is computed from the distributions of ideas over matter corpus, that is needed to arrange a website corpus containing annotated ideas and has high procedure price. As instances ar already extracted from matter corpus and annotated by ideas in KGs, graph-based IC is projected to reason IC supported the distributions of ideas over instances. Through experiments performed on acknowledge word similarity datasets, we have a tendency to show that the wpath linguistics similarity technique has created statistically important improvement over alternative linguistics similarity ways. Moreover, in an exceedingly real class classification analysis, the wpath technique has shown the most effective performance in terms of accuracy and F score.

Index Terms: Semantic Similarity, Semantic Relatedness, Information Content, Knowledge Graph, WordNet, DBpedia.

I. Introduction

With the speedy development of cloud computing, huge knowledge and public cloud services are wide used. The user will store his knowledge within the cloud service. though cloud computing brings nice convenience to enterprises and users, the cloud computing security has continually been a significant hazard. For users, it's necessary to require full advantage of cloud storage service, and additionally to confirm knowledge privacy. Therefore, we'd like to develop a good access management resolution. Since the standard access management strategy cannot effectively solve the safety issues that exist in knowledge sharing.

Data security problems brought by knowledge sharing have seriously hindered the event of cloud computing, varied solutions to attain coding and decoding of information sharing are projected. In 2007, Bethencourt et al. initial projected the ciphertext policy attribute-based coding (CP-ABE). However, this theme doesn't contemplate the revocation of access permissions. In 2011, Hur et al. advises a fine-grained revocation theme however it will simply cause key written agreement issue. Lewko et al. used multi authority ABE (MA-ABE) to unravel key written agreement issue. however the access policy isn't versatile. Li et al conferred knowledge sharing theme supported general attribute coding, that endows totally differentusers' different access rights. But it's not economical from the quality and potency. In 2014, Chen et al. projected Key-Aggregate coding formula, effectively shortening the length of the ciphertext and therefore the key, however just for things wherever the information owner is aware of the user's identity. These schemes higher than solely target one side of the analysis, and don't have a strict uniform standards either. during this paper, we tend to gift a additional systematic, versatile and economical access management theme.

To this finish, we tend to create the subsequent main contributions: one. we tend to propose a unique access system referred to as PSACS, that is privilege separation supported privacy protection. The system uses Key-Aggregate coding (KAE) theme and Hierarchy Attribute-based coding (HABE) theme to implement browse access management theme within the PSD and course severally. The KAE theme greatly improves access potency and therefore the HABE theme for the most part reduces the task of one authority and protects the privacy of user knowledge. 2. Compared with the MAH-ABE theme that doesn't ask the write access management, we tend to exploit Associate in Nursing Improved Attribute-based Signature (IABS) [7-9] theme to enforce write access management within the PSD. during this manner, the user will pass the cloud server's signature verification while not revealing the identity, and with success modify the file. 3. we offer a radical analysis of security and quality of our projected PS-ACS theme. The practicality and simulation results give knowledge security in acceptable performance impact, and prove the practicableness of the theme.

II. Related Works

Cloud computing is an emerging computing paradigm within which resources of the computing infrastructure area unit provided as services over the web. As promising because it is, this paradigm conjointly brings forth several new challenges for information security and access management once users source sensitive information for sharing on cloud servers, that aren't among an equivalent trusty domain as information homeowners. to stay sensitive user information confidential against untrusted servers, existing solutions sometimes apply scientific discipline strategies by revealing information decoding keys solely to licensed users. However, in doing thus, these solutions inevitably introduce a significant computation overhead on the information owner for key distribution and information management once fine-grained data access management is desired, and so don't scale well. the matter of at the same time achieving fine-grainedness, quantifiability, and information confidentiality of access management truly still remains unresolved. This paper addresses this difficult open issue by, on one hand, shaping and imposing access policies supported information attributes, and, on the opposite hand, permitting the information owner to delegate most of the computation tasks concerned in fine-grained information access management to untrusted cloud servers while not revealing the underlying data contents. we have a tendency to come through this goal by exploiting and unambiguously combining techniques of attribute-based cryptography (ABE), proxy re-encryption, and lazy re-encryption. Our planned theme conjointly has salient properties of user access privilege confidentiality and user secret key answerability. intensive analysis shows that our planned theme is extremely economical and incontrovertibly secure below existing security models.

Cloud computing has several issues. Cloud computing could be a promising computing paradigm that recently has drawn intensive attention from each world and trade. By combining a group of existing and new techniques from analysis areas adore Service -Oriented Architectures and virtualization, cloud computing is considered such a computing paradigm within which resources within the computing infrastructure area unit provided as services over the web. information security, because it exists in several alternative applications, is among these challenges that may raise nice issues from users once they store sensitive data on cloud servers. These issues originate from terribly the actual fact that cloud servers area unit sometimes operated by business suppliers that area unit very probably to be outside of the trusty domain of the users. information confidential against cloud servers is thus often desired once users source information for storage within the cloud. In some application systems, information confidentiality isn't solely a security/privacy issue, however conjointly of jural issues. what is more, we have a tendency to observe that there also area unitcases within which cloud users themselves are content suppliers. They publish information on cloud servers for sharing and want fine-grained information access management in terms of that user (data consumer) has the access privilege to that kinds of information. within the attention case, to Illustrate, a centre would be the information owner WHO stores variant attention records within the cloud. it'd permit information customers adore doctors, patients, researchers and etc., to access varied kinds of attention records below policies admitted by HIPAA. To enforce these access policies, the information homeowners on one hand would love to require advantage of the abundant resources that the cloud provides for potency and economy; on the opposite hand, they will wish to stay the information contents confidential against cloud servers. we have a tendency to address this open issue and propose a secure and climbable fine-grained information access management theme for cloud computing. Our planned theme is partly supported our observation that, in application eventualities every file will be related to a group of attributes that area unit important within the context of interest. The access structure of every user will so be outlined as a novel logical expression over these attributes to replicate the scope of information files that the user is allowed to access. because the logical expression will represent any desired file set, fine-graininess of information access management is achieved. To enforce these access structures, we have a tendency to outline a public key element for every attribute. information files area unit encrypted victimization public key parts admire their attributes. User secret keys area unit outlined to replicate their access structures so a user is in a position to decode a cipher text if and provided that the information file attributes satisfy his access structure. Such a style conjointly brings regarding the potency profit, as compared to previous works, in that, 1) the quality of cryptography is simply connected the quantity of attributes associated to the information file, and is freelance to the quantity of users within the system; and 2) file creation/deletion and new user grant operations simply have an effect on current file/user while not involving system-wide file update or re-keying. One very difficult issue with this style is that the implementation of user revocation, which might inevitably need re-encryption of information files accessible to the going away user, and should would like update of secret keys for all the remaining users. If of these tasks area unit performed by the information owner himself/herself, it'd introduce a significant computation overhead on him/her and should conjointly need the information owner to be continuously on-line. To resolve this difficult issue, our planned theme allows the information owner to delegate tasks of information file re-encryption and user secret key update to cloud servers while not revealing data contents or user access privilege information. we have a tendency to come through our style goals by exploiting a unique scientific discipline primitive, particularly key policy attribute-based cryptography.

2.1 Existing System

One of the drawbacks of typical knowledge-based approaches (e.g. path or lch) in addressing such task is that the linguistics similarity of any 2 ideas with constant path length is that the same (uniform distance problem).

we propose a we have a tendency toighted path length (wpath) technique to mix each path length and IC in measurement the linguistics similarity between ideas. The IC of 2 constructs' LCS is employed to weight their shortest path length so those concept pairs having same path length will have totally different|completely different} linguistics similarity score if they need different LCS.

III. PROPOSED SYSTEM

The wpath linguistics similarity methodology is to cipher each the structure of the thought taxonomy and also the applied mathematics data of ideas. what is more, so as to adapt corpus-based IC ways to structured KGs, graph primarily based IC is planned to figure IC supported the distribution of ideas over instances in KGs. Consequently, mistreatment the graph-based IC within the wpath linguistics similarity methodology will represent the specificity and hierarchical data structure of the ideas in a very kilogram.

This paper considers the matter of measurement linguistics similarity between ideas in KGs. the most contributions of this work is also summarized as below.

We propose a way for measurement the linguistics similarity between ideas in KGs.

We propose a way to figure IC supported the specificity of ideas in KGs.

We measure the planned ways in gold commonplace word similarity datasets.

We measure the linguistics similarity ways in side class classification.

IV. MODULES

4.1 Module Description:

In this project, Computing Semantic Similarity of Concepts, we have three modules.

- ❖ User module
- ❖ Multi-authorityAccess control
- ❖ Public cloud storage.

WPath Semantic Similarity Metric:

The knowledge-based linguistics similarity metrics mentioned within the previous section are in the main developed to quantify the degree to that 2 ideas are semantically similar exploitation data drawn from thought taxonomy or IC. Metrics take as input a combine of ideas, and come back a numerical price indicating their linguistics similarity. several applications suppose this similarity score to rank the similarity between totally different pairs of ideas.

Considering each benefits and drawbacks of standard knowledge-based linguistics similarity ways, we have a tendency to propose a weighted path length (wpath) technique to mix each path length and IC in measure the linguistics similarity between ideas. The IC of 2 thoughts' LCS is employed to weight their shortest path length so those concept pairs having same path length will have completely different linguistics similarity score if they need different LCS..

Graph-Based Information Content:

Conventional corpus-based IC needs to organize a site corpus for the thought taxonomy and so to work out IC from the domain corpus in offline. The inconvenience lies within the high procedure price and problem of getting ready a site corpus. additional specifically, so as to work out corpus-based IC, the ideas within the taxonomy ought to be mapped to the words within the domain corpus. Then the looks of ideas area unit counted and therefore the IC values for ideas area unit generated. during this approach, the extra domain corpus preparation and offline computation might forestall the appliance of these linguistics similarity ways looking forward to the IC values (e.g., res, lin, jcn, and wpath) to KGs, particularly once the domain corpus is too little or the kilo is often updated. Since KGs already well-mined structural data from matter corpus, we have a tendency to gift a convenient graph-based IC computation technique for computing the IC of ideas during a kilo supported the instance distributions over the thought taxonomy.

Word Similarity Evaluation:

All the datasets delineated higher than contain a listing of triples comprising 2 words and a similarity score denoting word similarity judged by human subjects. The human ratings on those word pairs are tested to be extremely replicable.

This indicates that human assessment regarding linguistics similarity between words is remarkably stable over an outsized time span and such datasets containing human ratings will be faithfully used for evaluating linguistics similarity strategies. Since those datasets contain totally different coverage of word pairs, we tend to use all the datasets for analysis so as to gift a additional completed and objective experiment.

Those datasets square measure used for evaluating word similarity. However, the linguistics similarity metrics conferred during this paper square measure used for ideas, instead of words. we tend to convert those concept-to-concept linguistics similarity metrics into a word-to-word similarity metrics by taking the maximal similarity score over all the ideas that square measure the senses of the words.

V. CONCLUSION

Measuring linguistics similarity of ideas could be a crucial element in several applications that has been conferred within the introduction. during this paper, we tend to propose wpath linguistics similarity technique combining path length with IC. the fundamental plan is to use the trail length between ideas to represent their distinction, whereas to use IC to contemplate the commonality between ideas. The experimental results show that the wpath technique has created statistically vital improvement over different linguistics similarity strategies. what is more, graph-based IC is planned to reckon IC supported the distributions of ideas over instances. it's been shown in experimental results that the graph-based IC is effective for the res, sculptor and wpath strategies and has similar performance because the typical corpus-based IC. Moreover, graph-based IC contains a variety of advantages, since it doesn't needs a corpus and permits on-line computing supported on the market KGs. supported the analysis of a straightforward side class classification task, the planned wpath technique has conjointly shown the simplest performance in terms of accuracy and F score.

In this paper, we tend to evaluated the planned technique within the word similarity dataset and straightforward classification victimization the foremost established analysis technique. additional analysis of linguistics similarity strategies in different applications considering the compartmentalization relation can be helpful and might be one in every of our future works. what is more, this paper in the main mentioned linguistics similarity instead of general linguistics connection. Therefore, another future work can be in finding out the mixture of knowledge-based strategies with the corpus-based strategies for linguistics connection. Finally, since we tend to combined WordNet and DBpedia along during this paper, we'd any explore victimization the planned approaches for mensuration the entity similarity and connection in KGs..

References

[1] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, “Freebase: a collaboratively created graph database for structuring human knowledge,” in Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM, 2008, pp. 1247–1250.

[2] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, “Dbpedia-a crystallization point for the web of data,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 7, no. 3, pp. 154 – 165, 2009, the Web of Data.

[3] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, “Yago2: A spatially and temporally enhanced knowledge base from wikipedia (extended abstract),” in Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, ser. IJCAI ’13. AAAI Press, 2013, pp. 3161–3165.

[4] I. Horrocks, “Ontologies and the semantic web,” Commun. ACM, vol. 51, no. 12, pp. 58–67, Dec. 2008. [Online]. Available: http://doi.acm.org/10.1145/1409360.1409377

[5] G. A. Miller, “Wordnet: a lexical database for english,” Communications of the ACM, vol. 38, no. 11, pp. 39–41, 1995.

[6] R. Navigli and S. P. Ponzetto, “Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network,” Artificial Intelligence, vol. 193, pp. 217–250, 2012.

[7] E. Hovy, R. Navigli, and S. P. Ponzetto, “Collaboratively built semi-structured content and artificial intelligence: The story so far,” Artificial Intelligence, vol. 194, pp. 2 – 27, 2013, artificial Intelligence, Wikipedia and Semi-Structured Resources.

[8] R. Navigli, “Word sense disambiguation: A survey,” ACM Computing Surveys (CSUR), vol. 41, no. 2, p. 10, 2009.

[9] A. Moro, A. Raganato, and R. Navigli, “Entity linking meets word sense disambiguation: a unified approach,” Transactions of the Association for Computational Linguistics, vol. 2, pp. 231–244, 2014.

[10] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum, “Kore: Keyphrase overlap relatedness for entity disambiguation,” in Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ser. CIKM ’12. New York, NY, USA: ACM, 2012, pp. 545–554.