Mitigating Relation Completion Problem in Big Data Application

Bhor Priyanka M.[1], Deolekar Shweta R.[2], Gaikwad Revati R.[3], Gonjare Shraddha M.[4]

Prof. Y. S. Deshmukh.
Department of Information Technology,
SRES COE, Maharashtra, India

ABSTRACT

Relation completion (RC) as one recurring problem in big data applications such as Entity Reconstruction and Data Enrichment. Given a semantic relation R, RC attempts at linking entity pairs between two entity lists under the relation R. To accomplish the RC goals, It formulate search queries for each query entity based on some auxiliary information, so that to detect its target entity from the set of retrieved documents. For instance, a pattern-based method (PaRE) uses extracted patterns as the auxiliary information in formulating search queries. However, high-quality patterns may decrease the probability of finding suitable target entities. As an alternative, we find CoRE method that uses context terms learned surrounding the expression of a relation as the auxiliary information in formulating queries. The experimental results based on several real-world web data collections demonstrate that CoRE reaches a much higher accuracy than PaRE for the purpose of RC (Relation completion).

Keyword:- *Context-aware relation extraction, relation completion, relation query expansion, information extraction.*

1. INTRODUCTION

The Big Data is large amount of data, giving rise to a new generation of applications that attempt at linking Related data from different sources. This data is typically unstructured and naturally lacks any binding information. Also it refers as data set or combination of data set whose size or rate of growth make them difficult to be captured. Linking this data clearly goes beyond the capabilities of current data integration systems. Today to get relevant and exact data is the major challenge. There are so many approaches to retrieve or to extract the relevant data like Pattern based System etc. There are some information extraction (IE) tasks such as named entity recognition (NER) and relation extraction (RE) which have been used to enable some of the emerging data linking applications such as entity reconstruction and data enrichment. so in this method relation completion (RC) as one of the recurring problem was occurs.

For a given semantic relation R, RC attempts at linking entity pairs between two entity lists under the relation R, for each query entity from a Query List, find its target entity from a Target List. This approach has some drawbacks as the number of retrieved documents, processing them incurs a large overhead and, those documents would include significant amount of noise. The Pattern- based semi-supervised Relation Extraction method (PaRE) is used for the purpose of RC, which is able to extract patterns of the relation R from the web documents. However, it relies on high-quality patterns which may decrease the probability of finding suitable target entities. The context aware relation extraction (CoRE) method is particularly designed for the RC task which overcomes the drawback of PaRE. In CoRE instead of representing a relation in the form of strict high-quality patterns, it uses context terms, which is called as relation-context terms (RelTerms) [1].

2. LITERATURE SURVEY

Zhixu Li, Mohamed A. Sharaf, Laurianne Sitbon, Xiaoyong Du, and Xiaofang Zhou [1] In this research paper authors identified relation completion (RC) is one recurring problem that is central to the success of novel big data applications. For given a semantic relation R, RC attempts at linking entity pairs between two entity lists under the relation R. This task was accomplish by RC goals, proposing to formulate search queries for each query entity based on some auxiliary information, so that it detect its target entity from the set of retrieved documents. For instance, a pattern-based method (PaRE) uses extracted patterns as the auxiliary information in formulating search queries. However, high-quality patterns may decrease the probability of finding suitable target entities. As an alternative, author proposed CoRE method which uses context terms learned surrounding the expression of a relation as the auxiliary information in formulating queries. The experimental results based on several real-world web data collections demonstrate that CoRE reaches a much higher accuracy than PaRE for the purpose of RC.

The authors Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr. and Tom M. Mitchell[2] has proposed in this paper architecture for a never-ending language learning agent, and also described a partial implementation of that architecture which uses four subsystem components that learn to extract knowledge in complimentary ways. After running for some days, this implementation populated a knowledge base with over some facts with an estimated precision. These results was illustrate that the benefits of using a diverse set of knowledge extraction methods which were amenable to learning, and a knowledge base also allows the storage of candidate facts as well as confident beliefs.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead and Oren Etzioni[3] In this paper the authors introduce Open IE from the Web, an unsupervised extraction paradigm that eschews relation-specific extraction in favor of a single extraction pass over the corpus during which relations of interest were automatically discover and efficiently stored. Unlike traditional IE systems that repeatedly incur the cost of corpus analysis with the naming of each new relation, Open IEs was one-time relation discovery procedure that was allows a user to name and explore relationships at interactive speeds. The paper also introduced TEXTRUNNER, which was a fully implemented Open IE system, and demonstrated its ability to extract massive amounts of high-quality information from a million Web page corpus. The system would also benefit from the ability to learn the types of entities commonly taken by relations.

Ravi Gummadi ,Anupam Khulbe , Aravind Kalavagattu ,Sanil Salvi, Subbarao Kambhampati [4] authors presented a unified approach that supports intelligent retrieval over fragmented web databases by mining and using inter-table dependencies. The authors also stated that experimental results demonstrated in this approach used by SMARTINT was able to strike a better balance between precision and recall than can be achieved by relying on single table or employing direct joins.

YuanhuaLvCheng Xiang Zhai [5] the authors introduced Pseudo-relevance feedback is an effective technique for improving retrieval results. Traditional feedback algorithms use a whole feedback document as a unit to extract words for query expansion, which is not optimal as a document may cover several different topics and thus contain much irrelevant Information. In this paper, authors give how to effectively select from feedback documents those words that are focused on the query topic based on positions of terms in feedback documents. They proposed a positional relevance model (PRM) to address this problem in a unified probabilistic way. The PRM is an extension of the relevance model to exploit term positions and proximity so as to assign more weights to words closer to query words based on the intuition that words closer to query words are more likely to be related to the query topic. They were developing two methods to estimate PRM based on different sampling processes. Experiment results on two large retrieval data sets show that the proposed PRM is effective and robust for pseudo-relevance feedback, significantly outperforming the relevance model in both document-based feedback and passage-based feedback.

Truc-Vien T. Nguyen and Alessandro Moschitti [6] the authors extend distant supervision (DS) based on Wikipedia for Relation Extraction (RE) by considering two things:

1. relations defined in external repositories, e.g. YAGO, and
2. any subset of Wikipedia documents

It shows that training data constituted by sentences containing pairs of named entities in target relations is enough to produce reliable supervision. These experiments with state-of-the-art-relation extraction models, trained on the above data, test set which highly improves the state-of-art in RE using DS. Additionally, these end-to-end experiments demonstrated that the extractors can be applied to any general text document.

3. PROPOSED SYSTEM

A. Problem Statement:

We identify relation completion (RC) as one recurring problem which deals with some applications such as Entity Reconstruction and Data Enrichment. This RC problem create in PaRE to overcome it CoRE is suggested to reduce it.

B. Proposed System Architecture:

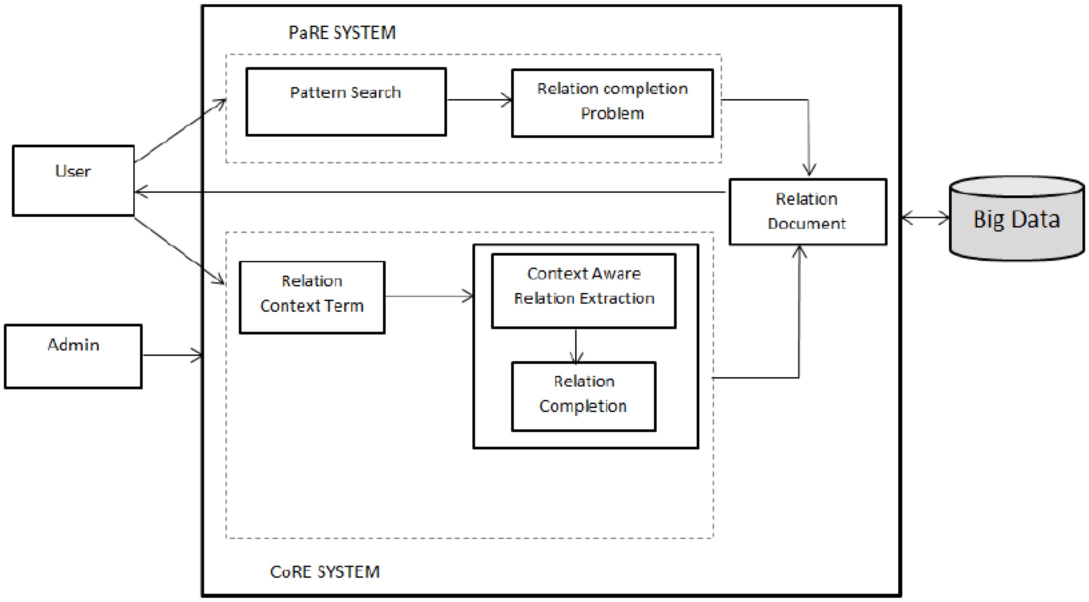

Figure 1: System Architecture

1. Pattern Based Search:

In the proposed scheme to provide secure data storage three participants are involved. These participants are Data Multiple RelQueries are posed, each of which is based on the query entity a in conjunction with one of the patterns extracted by the PaRE method (e.g., Brown joined + in), (Bob Brown works at), etc.). Using patterns as auxiliary information will generate very strict RelQueries, which will return the least number of web documents, but most of which are RelDocs. Hence, if a query entity a happened to appear in a webpage under one of the used Patterns, it will be quickly matched with its correct target entity. However, such assumption is unrealistic for many query entities that appear in very few web pages (i.e., long tail). For those entities, no web pages will be returned and will remain unmatched. This formulation is orthogonal to the pattern-based one above, where Multiple RelQueries are posed, each of which is based on the query entity α and an target entity β from the target list. Hence, each of the retrieved documents is processed to detect any of the patterns extracted by the PaRE method to justify whether (α, β) ϵ R. Obviously, this formulation incurs a large overhead as it requires posing a large number of RelQueries for each query entity as well as processing the documents retrieved by those queries.

2. Real Term Search:

CoRE utilizes the existing set of linked pairs towards learning Relation Expansion Terms (i.e., RelTerms) for any relation R. This task involves two main challenges:

I. Learning a set of high-quality candidate RelTerms from each existing linked pair.
II. Consolidating and pruning those individual candidate sets into a minimal global set of RelTerms that are used in the formulation of RelQueries.

CoRE formulates and issues a set of Relation Queries (i.e., RelQueries) for each query entity based on the set of learned RelTerms. However, there are many possible formulations, each of which is based on and a conjunction of RelTerms. Clearly, formulating and issuing all those queries will incur a large overhead, which is impractical. Hence, one major challenge is to minimize the number of issued RelQueries while at the same time maintaining high accuracy for the RC task. Towards achieving that goal, we propose one orthogonal technique is a confidence-aware termination (CA-Term) condition, which estimates the confidence that a candidate target entity is the correct one a tree-based query formulation method, which selects a small subset of RelQueries to be issued as well as schedules the order of issuing those RelQueries.

3. Context aware relation extraction:

This formulation is based on our proposed CoRE, in which multiple RelQueries are posed, each of which is based on the query entity in conjunction with several RelTerms extracted by the CoRE method (e.g., (“Bob Brown”+ “Department”), (“Bob Brown” +“Faculty”), etc.). By using RelTerms, a limited number of documents are retrieved, among which some are RelDocs that contain the correct target entity. Context-based formulation tries to strike a fine balance between a very strict RelQuery formulation (i.e., pattern based) and a very relaxed one (i.e., query-based). Towards this, CoRE exploits RelTerm towards a flexible query formulation in which a RelQuery is formulated based on the query entity α in conjunction with one or more RelTerms.

4. Relation Completion:

RelTerms from each of the existing individual linked pair, CoRE selects a set of general RelTerms from those candidates. The goal is to select a set of high-quality RelTerms for effective query formulation, and in turn accurate relation completion (i.e., finding target entities). In CoRE, this task takes place in two steps:

I. CoRE uses a local pruning strategy to eliminate the least effective RelTerms,
II. CoRE uses a global selection strategy to choose the most effective RelTerms.

During the local pruning step, CoRE verifies the effectiveness of each RelTerm in extracting the target entity for the linked pair from which it was learned. In particular, in the verification of a linked pair such as considered as a seed RelQuery without auxiliary information and each learned RelTerm used as a candidate to such seed query with auxiliary information.

4. PROPOSED MECHANISM:

Here a mechanism is proposed to provide secure data storage in Mobile Cloud Computing. This proposal uses the concept of Hash function along with several cryptographic tools to provide better security to the data stored on the mobile cloud. Here we also have a Trusted Third Party Auditor (TPA) who is very well trusted. TPA checks the integrity of the data stored on mobile cloud on behalf of the data owner. TPA checks the hash and message to verify the integrity of the data. In this scheme data owner has two keys, one of which is only known to him called private key and another is public key. Here message/_le is encrypted twice firstly, by owner's private key and secondly by public key of TPA. So this provides the confidentiality to the data of mobile user. In proposed method RSA algorithm is used for performing encryption and decryption which provides message authentication. Here the hash function of the message is also calculated to provide security to the data.

1. Key Generation:

Data Owner uses RSA algorithm for generation of public key and private key for himself. TPA also uses RSA algorithm for key generation. The private key of TPA is pk1 and of Data Owner is pk2, while public key of TPA is dl and public key of data owner is d2.

2. Key Sharing:

Key set of TPA: {pk1, dl}
Key set of DO: {pk2, d2}

3. Encryption:

Firstly, At first, data owner encrypt the message/ file (F) using his public key (d2) E(F,d2) and then generate the hash of encrypted message H(E(F,d2)). Then, the encrypted file is re-encrypted with public key (dl) of TPA E(E(F,d2),dl). After that the hash is re-encrypted with public key of TPA (dl) E(H(E(F,d2)),dl). Now, these two packages are appended and the result E(E(F,d2),dl) II E(H(E(F,d2)),dl) is sent to TPA. The encrypted Hash function of the message is stored by TPA to ensure the data integrity. TPA decrypts the package E(E(F,d2),dl) received, by its private key. TPA generates a random key for performing encryption on the message E(F,d2) generated after encryption. TPA uses DES (Data Encryption Standard) for performing encryption to provide better security. This generated random key is stored by TPA for performing decryption in future. The result is send to the cloud for storage.

4. Decryption:

When required to verify the data correctness, the encrypted package {Encrypt(E(F ,d2))} after DES operation stored on cloud is send to TPA. TPA firstly decrypts the message by random key stored by him. Then TPA generates the Hash of the encrypted file. Now, TPA decrypts the hash value stored by it, this decrypted hash value is compared with the one generated by it. Then according to the result obtained TPA sends file to owner indicating the correctness or not and the requested file. Here the file transferred to owner is encrypted by his public key so that only owner can decrypt it. Owner after receiving encrypted file, decrypt it by private key of himself.

Computational Overhead:

ENCRYPTION PROCESS			
	Exponential Operations	Hash Function	Pairing Operation
USER	3	1	1
TPA	1	0	1

	DECRYPTION PROCESS		
	Exponential Operations	Hash Function	Pairing Operation
USER	1	0	0
TPA	2	1	0

Figure 2: Computational Overhead

Storage Requirement:

PARTICIPANT	**STORAGE REQUIREMENT**
Mobile Device	Stores Public key of TPA and Public and Private key of owner
TPA	Stores Public key and private key of itself and Hash of the file received from mobile user
CSP	Stores encrypted file of mobile user

Figure 3: Storage Requirement

5. EXPERIMENTAL RESULT

An application has been designed and developed in android platform where client have an android application from which the client can upload the Text, Image, etc. on the cloud in encrypted format and retrieve the same and decrypt it using his Private Key for original file. Therefore the required result has been achieved.

IJARIIE

6. CONCLUSION

In this Paper, we conclude that the result get from PaRE method is some time not accurate because it uses particular pattern search for extracting the information. So we conclude that CoRE is effective and efficient method which uses RelTerm instead of pattern for extraction of information from this we get the accurate result for which we fire a query. Also the RC task is solved by CoRE method, efficiently and accurate than PaRE method. We also demonstrate the effectiveness and efficiency of our proposed techniques in learning relation terms and formulating search queries.

REFERENCES

1. Preeti Garg, Dr. Vineet Sharma "An Efficient and Secure Data Storage in MobileCloud Computing through RSA and Hash Function" International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), (IEEE-2014) pp.334-339.

2. Lifei Wei, Haojin Zhu, Zhenfu Cao, Xiaolei Dong, Weiwei Jia, Yunlu Chen, Athanasios V. Vasilakos "Security and privacy for storage and computation in cloud computing" (ELSEVIER 2014),Information Sciences 258 (2014) pp.371-386.

3. Mazhar Ali, Samee U. Khan, Athanasios V. Vasilakos "Security in cloud computing: Opportunities and challenges" (ELSEVIER 2014), Information Sciences 305 (2015)pp.357-383.

4. Sujithra M, Padmavathi G, Sathya Narayanan "Mobile Device Data Security: A Cryptographic Approach by outsourcing Mobile data to cloud" Procedia Computer Science47 (2015) pp.480-485.

5. M Sulochana, Ojaswani Dubey "Preserving Data Confidentiality using Multi-CloudArchitecture “Procedia Computer Science 50 (2015) pp.357-362.

6. M. Thangavel, P. Varalakshmi, Mukund Murrali, K. Nithya "An Enhanced and Secured RSA Key Generation Scheme (ESRKGS)" journal of information security andapplications 20 (2015) pp.3-10.

7. Miss. Shakeeba S. Khan, Miss. Sakshi S. Deshmukh "Security in Cloud ComputingUsing Cryptographic Algorithms" IJCSMC, Vol.3, Issue.9, September 2014, pp.517-525.

IJARIIE

BIOGRAPHIES

	Patil Rahul V. is pursuing B.E. Information Technology in SRES COE, Kopargaon. His area of research interest include Cloud computing.
	Shinde Pratik M. is pursuing B.E. Information Technology in SRES COE, Kopargaon. His area of research interest include Cloud computing.
	Sonawane Shriraj M. is pursuing B.E. Information Technology in SRES COE, Kopargaon. His area of research interest include Cloud computing.
	Somwanshi Akash B. is pursuing B.E. Information Technology in SRES COE, Kopargaon. His area of research interest include Cloud computing.