ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

A Novel Document Mapping Framework Using Semantic Structural Summary

Girisan.E.K[1], Shareeja.T[2]

Assistant Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India [1]

M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India [2]

ABSTRACT: With the huge development of internet, the web document management became tough and defective. Web data extraction and management using Graph mining and ontology is the main aim of this proposal. Ontology is created with the suitable entities and their relationships and is considered in resume extraction and filtering domain. Every document i.e. resume is split into four different categories. Attribute values are extracted from the resume documents. These values are updated in four different Resource Description Framework (RDF) files (nonSQL) for each resume through ontology mapping. Resumes are ranked based on Jaccard index and then the ontology is updated correspondingly. Thus a novel web management system is proposed which creates RDF files and also contributes a highly effective Semantic Structural Summary Based technique. This consists of querying, document search or subgraph search, and content summarization process using Probabilistic feature filtering algorithm.

KEYWORDS: Graph Mining, Information Retrieval, Semantic, Ontology, Resume Ontology, RDF, Document retrieval

I. INTRODUCTION

Graph mining is a popular area of research in the current era. This popularity gained due to its numerous application areas such as computational biology, software bug localization and computer networking. Besides these application areas, there are numerous categories of data like XML and other semi structured data that can typically be represented as graphs. Data representation and data management have become a tedious process in the recent data management [1]. Every domain and application follows different types of data and different structured information. Making them into a structured one is the main challenge. So graph mining is one of the processes of mining structured data for deep analysis. Graph mining contains various algorithms, and the requirement of applications is many and varied; it is not similar to one another. Hence, the graph mining algorithms are different for each domain and application. Graph mining is categorized according to two types of data- one is Web data and another one is XML data.

Web Data: In the web data, the node count is massive and the number of edges is also huge. It is associated with the massive domain issue. Such a domain needs a complete design or framework to summarize and represent the graph dataset into a structured summarized form [2].

XML data: The other type of domain data is XML (eXtensible Markup Language) data, which is a commonly known general data structure. The XML data can be represented with labels at the time of summarization. These XML data are often considered as independent data in the graph mining field [3].

Yet another form of XML data is referred as RDF **(Resource Description Framework)** files. These files contain all Meta descriptions [4]. In this proposal, RDF resume data are created and summarized with the ability to represent property values that exist but are unknown or partially known using constraints. We also propose a graph based data framework for storing and organizing resume documents. This includes the ontology mapping process. It is used to retrieve the relevant resume information from the RDF files.The research proposal aims at performing a fast and accurate resume search from the RDF document. The implementation of ontology and semantic concepts in the RDF data extraction helps to identify the exact data needed from the RDF files. The proposed work aims to develop a new indexing scheme for effective data identification and retrieval from the RDF, which is a non SQL document.

DOI: 10.15680/IJIRCCE.2016. 0409138

II. RELATED WORK

Several traditional mining applications used graph data. In such applications, the mining of graph data is a difficult and challenging one. To overcome the challenges and troubles of graph data mining, numerous techniques and algorithms are proposed. The types of algorithms are identified as clustering, classification and frequent pattern mining algorithms.

In [5] Mary am Fazel-Zarandi1, Mark S. Fox2 presented an approach using graph mining in the resume management application. This was developed to perform job matching process using matchmaking model, which is based on the description logics and similarity model. It is an ontology driven hybrid approach to effectively match job seekers and job advertisements based on the relationship. The approach uses a deductive model to determine the kind of match between a job seeker and an advertisement, and applies a similarity based approach to rank applicants. The main drawback of this approach is that it doesn't provide the automatic discovery process. With the use of domain ontology, this issue can be rectified. But the author left this process for future work. The authors specified that ontology can be used to formally define the semantics of information sources, dependencies between data, relationships between information sources and experts, and trust relationships to improve recognition and extraction.

In [6].Ujjal Marjit, Kumar Sharma and Utpal Biswas presented Linked Data approach to discover and aggregate resume information into the structured data. Using Linked Data technology, the data dependency can be detected. The discovery of data after successful aggregation may also resolve the heterogeneity issues in the resume management process. It helped to discover resume information by enabling the task aggregation and sharing and reusing of the information among different resume information providers and organizations. The data linked technology proposed by these authors was also not sufficient because, the authors failed to perform the domain linkage concepts. So this approach needed several manual discovery verification steps which decreased the efficiency of the system.

In [7].Kopparapu S.K described a system for automated resume information extraction to support fast resume search and organization. Automated resume information extraction process is capable of obtaining several important informative fields from an unstructured resume using a set of natural language processing (NLP) techniques, which comes under the text mining process. This technique performed the automatic resume management and improved the existing problem in [6]. The system is capable of extracting six major fields of information as defined by HR-XML. Finally this method yields 91 % and 88% precision and recollection respectively.

In [8] Celik Duygu, Karakas Askyn, Bal Gulsen, Gultunca Cem, proposed ontology-based information extraction system. This extraction process is called Ontology based Resume Parser (ORP). This ORP converts the resumes into ontological format using concept matching method. The overall usage of the ORP system is based on concept matching and ontological rules for English and Turkish resumes. It also provides semantic analysis of data and parses related information. Moreover, it is based on the Ontology Knowledge Base (OKB) method which transfers plaintext resume into ontology form and performs the inference. From the above literature, we found several resume management applications are developed using text and graph mining approaches. However, the resume data management is much complicated if the data's are unstructured. The semi-structured and un-structured data handling processes are maximum handled by the ontology process. So the proposed system need more effective tool to handles the problem from above previous works.

III. PROBLEM DEFINITION

Unstructured data extraction and management in the high dimensional environment is a tedious and tough process. The current resume management from several sources creates such issue. The conversion of such documents into a manageable format is the main aim of the proposed system. Mining graph data and summarizing with the semantic structure is a major issue. The studies in the literature show that the ontology framework gives a better result in the RDF data management. But the retrieval and ontology construction process should not be performed every time. So by limiting the number of iterations and the amount of time consumed the efficiency of the proposed system can be decreased. Considering the above, the proposed system designed a new RDF management tool with effective data mining techniques.

IV. PROPOSED SYSTEM

The unstructured data management is a complicated one. However the data management is performed by different types of tools and techniques. But RDF is one of the recent and popular file formats. RDF data model presents unique challenges to efficient storage for the high dimensional dataset, indexing and querying applications. The ever increasing number of job opportunities over web creates numerous applications and data related to the job. One kind of application related to the job seeking process is the resume management, which handles n number of resumes from different sources and different formats. Some tools are available in the web to find and manage the resumes. But the management of unstructured resume is still a problematic. In this paper, we propose a model for extracting resume information from different unstructured sources and make them easy to manage. The extracted resumes are divided into four parts which are personal details, skill details, and work and education details. These data are converted into corresponding Resource Description Framework (RDF) files for each resume through ontology framework. After the conversion, the resumes are ranked using **Jaccard index for** fast data retrieval and mapping. The proposed system has the following contributions towards the above issue.

- The proposed system collects the unstructured resume documents from the web. This type of extraction uses the basic API based crawling process from different domains.
- The next process is converting the data into the RDF format, which is a non SQL format. This conversion made by applying the ontology process.
- We contribute a new **Semantic Structural Summary Based Approach** for creating the resume summary and performing the search. This helps to find the documents from the RDF based on the semantic analysis.
- The collected documents are ranked and indexed using Jaccard index method, which is a well known similarity finding and indexing method. The main advantage of using this method is that it finds the coefficient of relationship between two unstructured free style data.
- Finally the Probabilistic Feature Filtering Algorithm (PFF) is introduced to filter resume documents effectively with unstructured query.

A. DATA CRAWLING AND ONTOLOGY CONSTRUCTION PROCESS:

The data crawling process is the extraction of resume documents from different sources. Before converting into the RDF files, the system should perform ontology construction process. The following steps are involved in the ontology construction process.

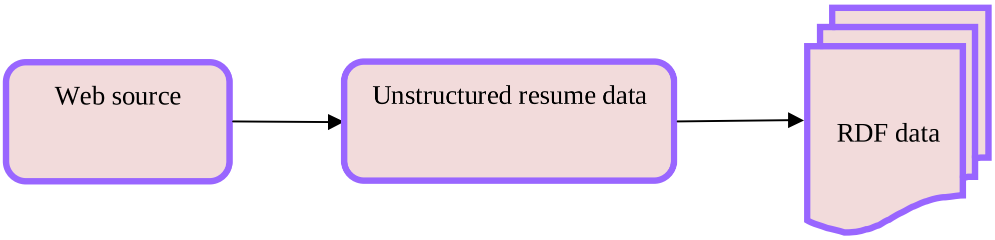


Fig 1.0 the data crawling and RDF conversion process

Fig 1.0 shows the first step of implementation, which retrieves resumes from the web sources as unstructured free style resume data and will be converted into the RDF format after the construction of ontology.

a. **RDF files:** Resource Description Framework (RDF) files are self-describing data model. It means that the schema information is embedded with the data are unspecified. And it follows a graph based architecture, where the prior structure can be unspecified. it gives a lot of flexibility to manage any structured or un-structured data and deal with changes in the data's structure seamlessly at the application level.

b. **Ontology:** Ontology is defines as an explicit specification of conceptualization, it captures the structure of the domain and describes the knowledge form that. Its defines the relationship between two or more

DOI: 10.15680/IJIRCCE.2016. 0409138

entities and interactions in some particular domain of knowledge or practices. In the proposed system, the ontology framework gives the knowledge based results of given dataset.

Computing important terms: The first step of ontology construction is the process of finding imported terms and words from the resumes. This step includes the detection of important and required terms in the wish list by analyzing the nouns and verbs.

Define concept taxonomies: The next process is to classify the concepts in a hierarchy. In this process, not all concepts will own a hierarchical structure. This step uses both top-down and bottom–up approaches.

Define relations: After defining the concept taxonomies, the system performs the relation identification process. In this paper the relations are represented using the semantic structure. For example, the terms education and courses are similar, and also the knowledge in computer languages such as C, C++ etc of a user may be related with one another.

Define attributes: At this step, some of the terms listed are considered to define the attributes. The attributes are constructed after the correlation detection between two or more entities. After successful detection of attributes the complete ontology tree will be constructed.

Define instances: The next step after attribute selection is finding instances under each attribute. In some cases, more than one instance may be present. Such data can be represented as a sub hierarchy instance in the ontology process. This helps to describe all the relevant instances. These will appear as having a name, a concept to which they are related and having unique attribute names and values.

B. INDEXING AND QUERY PROCESSING TECHNIQUES:

For effective data retrieval, the Jaccard indexing is proposed. The Jaccard index is a technique to find the coefficient of similarity between two resumes. It considers the resumes and indexes them based on the Jaccard distance. Thus, the data in the resume will be organized effectively. If $X=(x_1,x_2,x_3\ldots.x_n)$ and $Y=(y_1,y_2,y_3\ldots.y_n)$ are two vectors with all real xi,yi ≥ 0, then their Jaccard similarity coefficient is defined as specified below

$$J(\mathbf{x},\mathbf{y}) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)},$$

Equation 1.0

The equation 1.0 shows the Jaccard similarity co-efficient function, where x is considered as the Resume1 and y as Resume 2. All the features in x and y are defined as x_1, $x_2\ldots x_n$ and y_1, $y_2,\ldots y_n$ respectively. For every feature the minimum and maximum similarity features are mined. Finally the value of J(x,y) shows the similarity distance between x and y. After measuring the similarity between resume data, the system performs the data filtering process.

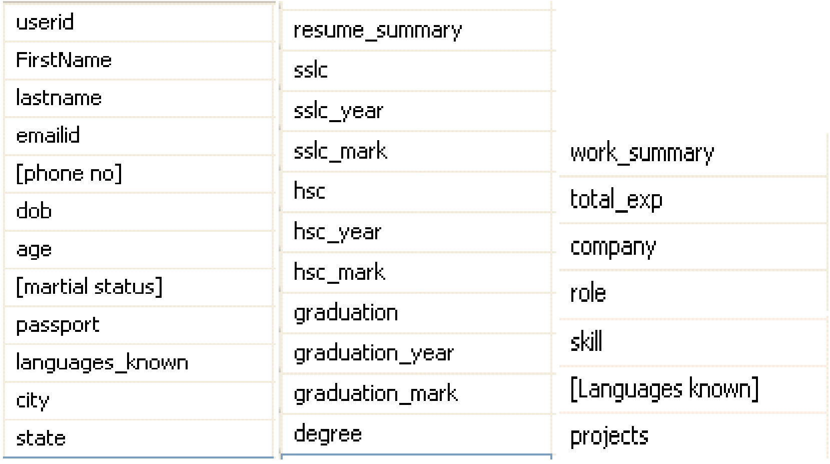
V. EXPERIMENTS AND RESULTS

A. DATASET FOR THE IMPLEMENTATION:

The system uses a dynamic synthetic dataset, which can be any number of user resumes that can be crawled from different job searching sites such as Monster.com, Linked in etc. Data collection is the first step proposed work, so n number of resume documents are collected and used in the proposed system. Many details in the resumes which are relevant for the comparison and evaluation are analyzed and sort out. The experiments are carried out in visual studio framework 4.5, and C#.net language. Initially the data has been collected from different sources and the data's are given as input to the application. The proposed system converts the different types of data into a single RDF file, and that will be stored in file storage.

The dataset for the experiment has n number of attributes, the attributes such as name, address, education and their work summary based details.

This contains the following attributes.



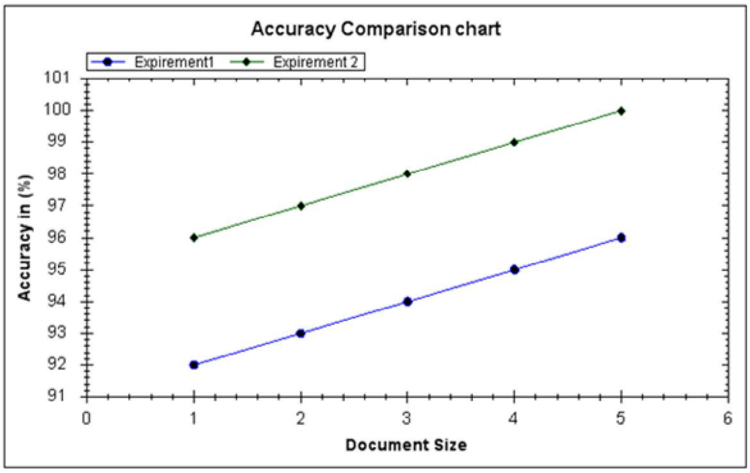
userid
FirstName
lastname
emailid
[phone no]
dob
age
[martial status]
passport
languages_known
city
state

resume_summary
sslc
sslc_year
sslc_mark
hsc
hsc_year
hsc_mark
graduation
graduation_year
graduation_mark
degree

work_summary
total_exp
company
role
skill
[Languages known]
projects

Table 1.0 attributes in the dataset

The table 1.0 shows the list of attributes used in the experiment. It presents the experimental results of the proposed system for over 50 datasets which are described below. The experiment is performed with two set of data structures, the experiment1 is conducted without using the index method for 50 records. The experiment 2 is applied in to the proposed framework with jaccard index. The both experiments are finally compared with two parameters; one is the accuracy in detection and the time taken for the document retrieval. The fig2.0 shows the experiments results for every 10 document size. When the documents are increasing, then the accuracy is also increases. These shows, for better and effective data management and semantic process, the data size should be higher than the average. If the document contains 30 or more records, the similarity can be detected and the retrieved results are verified effectively.



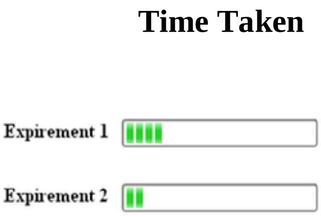


Fig 2.0 Comparison of accuracy between two experiments

The performance of the proposed schemes is evaluated on the basis of both time consumed and accuracy. Without loss of generality, this evaluates the Indexing delay and Retrieval Delay for deployed RDF files. Indexing delay indicates the time required to perform the coefficient calculation and indexing. Document retrieval delay indicates the time spent on processing on ontology mapping in the hierarchical form. The figure below shows the results and the comparison of the proposed system with the existing system.

Table2.0 Comparison table

The table 2.0 shows the performance comparison of the proposed method with other existing approaches based on the three different metrics Retrieval Delay, indexing delay and extraction accuracy. Performance comparison of proposed system using semantic onto creation with existing approaches based on Retrieval Delay

Parameters	Existing	Proposed
Retrieval Delay	2984.06	2108.08
Extraction Accuracy	90	94
Indexing delay	70.68	49.69

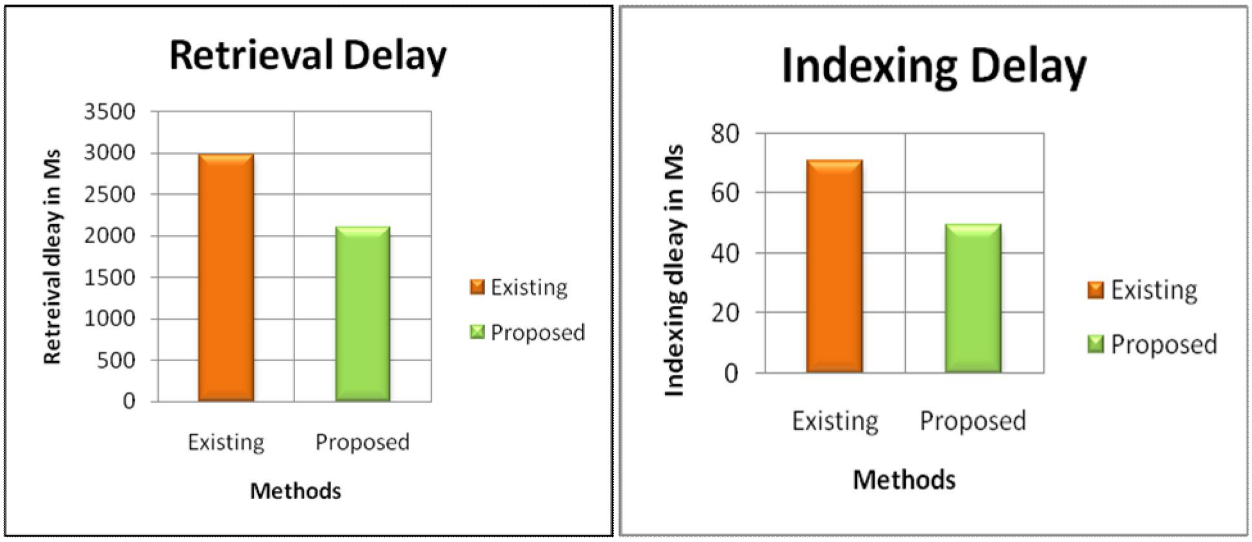


Fig 3.0 retrieval and indexing delay comparison chart

From the fig 3.0 it shows the performance measure based on the Retrieval Delay and the proposed approach took less time while comparing the other methods and the worst time complexity is existing system. The comparison is made with 100 data samples, the existing document retrieval method used static index process and Sql database query retrieval process, but the proposed system used RDF files with self structured ontology concept. The Jaccard index reduces the time and helps to retrieve the documents fast and accurate.

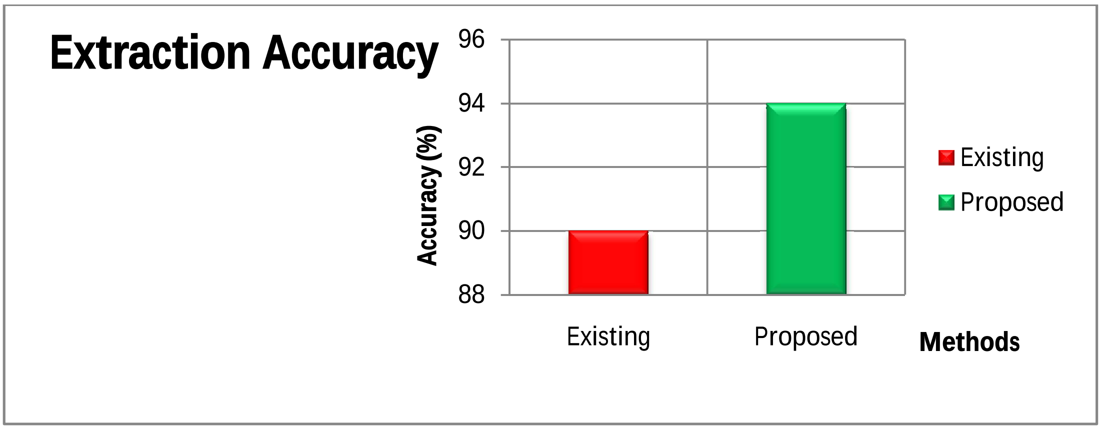


Fig 4.0 extraction accuracy comparison chart

The fig 4.0 shows the extraction accuracy comparison between existing and proposed systems. The accuracy is measured based on the correct number of documents extracted and wrongly detected documents. And the proposed approach is more accurate when comparing with the other methods.

VI. CONCLUSION

Data management from an unstructured source is a complicated process, where the RDF file format is self-describing model that can be used to handle such data’s in the application level. The proposed model collects resumes through web search and rank them based on Jaccard Index with semantic verification. In the proposed system Ontology plays the vital role while extracting and keeping the information relevant and updated in the RDF format. So, it reduces searching time for required information. The overall document search and indexing time is minimal when information is stored RDF files compared to other two models. The model also reduces the human effort required in seeking the relevant information.

REFERENCES

[1]. Gudes, Ehud. "Graph and web mining-motivation, applications and algorithms." *International Journal on Software Bug Management* (2010).

[2]. S. Abiteboul, P. Buneman, D. Suciu. Data on the web: from relations to semistructured data and XML. Morgan Kaufmann Publishers, Los Altos, CA 94022, USA, 1999

[3]. Li, Quanzhong, and Bongki Moon. "Indexing and querying XML data for regular path expressions." *VLDB*. Vol. 1. 2001.

[4]. Klyne, Graham, and Jeremy J. Carroll. "Resource description framework (RDF): Concepts and abstract syntax." (2006).

[5]. Maryam Fazel-Zarandi1, Mark S. Fox2, “Semantic Matchmaking for Job Recruitment: An Ontology-Based Hybrid Approach”, International Journal of Computer Applications (IJCA), 2013.

[6]. Ujjal Marjit, Kumar Sharma and Utpal Biswas, “Discovering Resume Information Using Linked Data”, in International Journal of Web & Semantic Technology, Vol.3, No.2, April 2012.

[7]. Kopparapu S.K, “Automatic Extraction of Usable Information from Unstructured Resumes to aid search”, IEEE International Conference on Progress in Informatics and Computing (PIC), Dec 2010.

[8]. Celik Duygu, Karakas Askyn, Bal Gulsen, Gultunca Cem, “Towards an Information Extraction System Based on Ontology to Match Resumes and Jobs”, IEEE 37th Annual Workshops on Computer Software and Applications Conference Workshops, July 2013.

DOI: 10.15680/IJIRCCE.2016. 0409138