CLUSTERING OF CATEGORICAL DATASET USING ONTOLOGY

R. Janani
Department of Computer Science, Alagappa University, India

Abstract

An important yet difficult topic is the incorporation of semantic knowledge from ontology into categorical documents. In addition, there are many subjects listed on the Internet based on computers. The aim of this system is to group documents depending on the nature of the data tested and to describe the techniques of testing in detail. It also reports on studies carried out to test the weighting scheme used to encode the relevance of concepts in papers. The approach employed the Table of Categorical Information before clustering the document to keep information on the idea before the concept was weighted. For the analysis, the system must employ ontology to describe and organise information and to cluster it from heterogeneous sources.

Keywords:

Ontology, Categorical Dataset, Semantic Relationship, Natural Language Processing

1. INTRODUCTION

The Internet booming includes a trillion textual documents on the World Wide Web. This reason placed an urgent demand for an ontology-built clustering approach to represent knowledge of a certain field throughout the World Wide Web [12]. In order to examine and use the vast number of text documents, a wide variety of ways have been created to help users navigate, summarise and arrange text papers effectively. But, as more text documents are being compiled, many systems have to rely on models such as the Semantic Web [9] as an emergency.

With the development of the Internet, document clusters have become a crucial technology. This also means that the quick, high-quality classification of documents plays a central role. The clustering or clustering of text is about semi-related groups being identified in an unstructured document collection [10].

Clustering has long been quite popular as it offers unique techniques to digest and generalise vast quantities of information [14]. The traditional clustering algorithms depend only on the duration and frequency of documentation that may readily be used in clusters. It also takes conceptual weight into account under the backing of ontology [12].

In the area of the Semantic Web, ontologies are now hot subjects. This system employs ontology concepts to construct a well-defined model of data with a good structure to effectively utilise this data and information [11].

The study focuses on the idea weighting and structural information tables, which can better identify the documents involved by taking advantage of the concepts of field ontology [15]. It is also crucial that huge amounts of text data are clustered in a trustworthy way [13]. This proposed method provides a novel way of compiling and clustering documents using ontology.

2. BACKGROUND THEORY

Text mining is a technology created by data mining for the analysis of textual data (free text, abstracts, etc), where the text is unclear and [1]-[3] apply. The overview of data mining is shown in Fig.1.

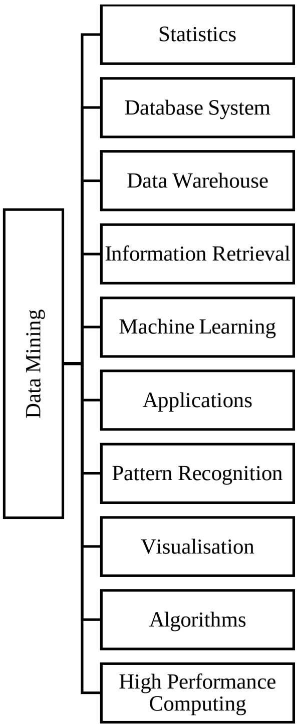

Fig.1. Overview of Data Mining

Ontology was traditionally characterised as the philosophy of what exists: the study of certain entity types in reality and the interplay between these entities. In the area of computer science and information technology, ontology defines a range of fundamental representatives for modelling a domain of knowledge or speech. The primitive representations are classes, attributes and relationships. In computer science, ontology is a technical term that refers to an item designed to facilitate knowledge modelling in certain fields, whether real or fictional. [4] - [8].

2.1 PROBLEM STATEMENTS

Digital community experts have recently noticed the huge growth in publications. Although search engines on the Internet provide scholars with an effective technique for searching for papers of interest, the vast amount of data remains an onerous endeavour.

One way to facilitate this is clustering, a practice employed in all disciplines. Ontologies can also help handle the search challenge, including research articles.

Most known methods of clustering text depend on the strength and frequency of documents using the TF-IDF form in the document only. However, this method just takes account of the times the words appear, while neglecting other aspects that may affect the term. This method is also just a binary form of weighting. This suggested system also takes account of the conceptual importance of picking the characteristics of documents with ontological support in order to make the use of ontology possible in the clustering process [16]. In addition, this system wished to use the ontology hierarchy structure before weighing phases, adding a category information table.

3. PROPOSED SYSTEM

The Fig.2 shows the system's principal development and offers a full description of the entire system process.

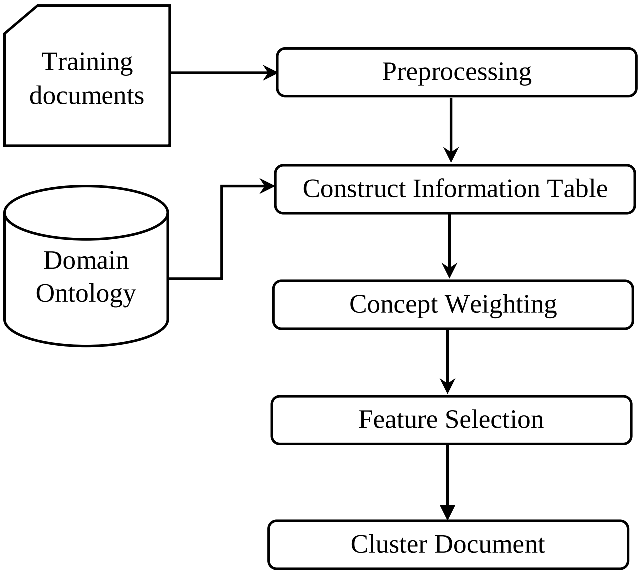

Fig.2. Overview of the System

The system is implemented in four sections. The first portion is ontology, the second part is an information table. This section provides the calculation of the weighting, and part final is the classification.

The objective of this project is to establish a domain-specific ontological framework that is used for clustering technology. This section provides a full description of the work of the systems that can be employed together with ontological ideas. The work was based on Google Search Engine information collected and inferred concerning thesis papers in the fields of image processing, the distribution system and the natural language processing field.

In view of the increasing demand of the Image Research, Distributed Systems, and Natural Language Processing communities, the concept of hierarchical data must be captured to provide an efficient means and model for those research fields. The system thereby produces an ontology to gain knowledge and has been designed for this area of inquiry. The fundamental steps in ontology development are simple. In text texts, the system has explored ontology building as illustrated in Fig.3. This ontology is captured by contemporary ontology editors, etc. in the OWL DL language. The design of the ontology will be influenced by the ontology system tasks.

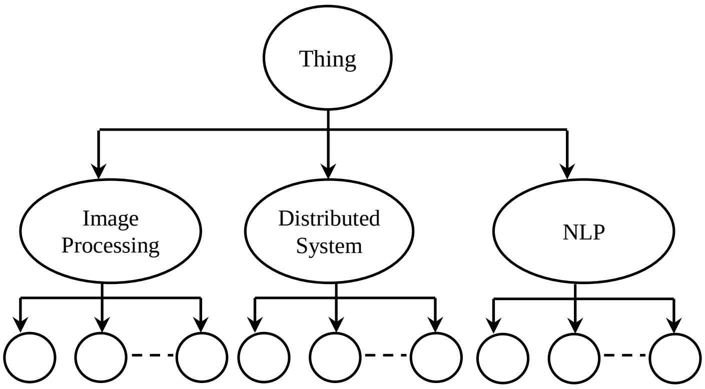

Fig.3. Part of Hierarchy Structure of the System

3.1 PRE-PROCESSING PHASE

The first stage of this phase is the collection of text documents. The document shall be moved to a format appropriate for the representation procedure at the pre-processing stage. In a large variety of machine formats, textual information is stored, such as PDF, DOC, PostScript, HTML, and XML. However, many documents are still stored in a simple pdf format. When the text document is collected by Google's search engine, it will electively be kept in the text file in the abstract of the article. After this stage, the machine will delete the stop words and the text content that has been extracted. High frequency words with no information are stopwords.

3.2 CONSTRUCT CATEGORICAL INFORMATION TABLE PHASE

Construct a Categorical Information Table for rapid reference to ontological information. The Categorical information table consists of the possibility of occurring with the extraction of the concept from ontology of a certain attribute value. Values are utilised before weighting the concept in the Categorical Information Table. Thus, a key sub-set of semantic traits that reflect text documents can be identified. Empirical results indicate that the selective selection matrix of features can be reduced by 90 percent or more by using key semantic characteristics for clustering and that clusters still provide the primary themes in text content.

Table.1. Construction Information Table

Attribute from ontology	Occurrence Correlation
Image………FingurePrint	1
Image…....................Scanner	1
Distributed Computing…Mobile	1
NLP……………………..N-Gram	1

3.3 CONCEPT WEIGHTING PHASE

The system calculates the weight in this step as indicated below [8].

$$W = L \times F \times CC + PoC \tag{1}$$

where

W is the weight of keywords,

L is the depth of concept in the ontology.

F is the times which count the words appear in the document, and if the concept is in the ontology

CC is taken as 1 and otherwise 0.

PoC is based on the probability of the concept in the document.

3.4 DOCUMENT CLUSTERING PHASE

One objective of this thesis is to group text materials based on their conceptual weight instead of keywords. This stage focuses on the incorporation into the clustering methods of the concept of semantic characteristics. The focus is on managing numerical and categorical data through clustering text methods. Textual information, however, has increased in relevance in recent years. Proper processing of this type of information requires semantic data extraction tools. The concept of the weighting concept was introduced to formally outline clustering documentation in the work of the system. The knowledge available is ontologically formalised. Clustering approaches that rely on ontology totally or partially. In the system, values weight rather than just phrase weight are concepts. Consequently, it should also benefit from better identification than non-semantic clusters by employing the results achieved in the first part of this research in the clustering procedures. On the other hand, a way to include the weight of semantic functions in an uncontrolled clustering algorithm was developed.

4. EXPERIMENTS

Three test cases were tested for the proposed system. These experiments were conducted using Google-downloaded publications. 800 documents have been downloaded from the Google Search page of the new World Wide Web Conference.

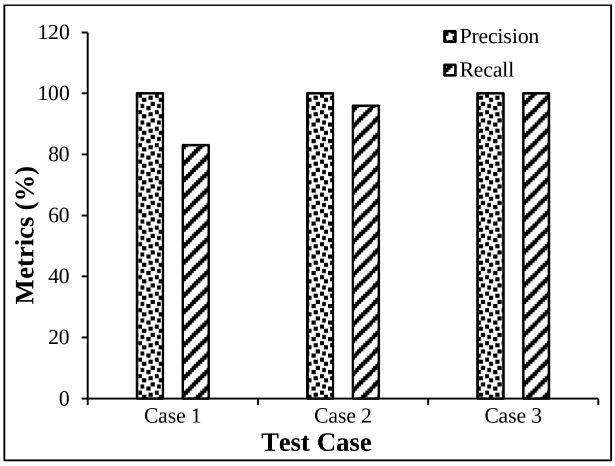

Fig.4. Testing data

Initially, the ontology construction and terminology training documents were collected and downloaded for five cents. 300 other documents were acquired as a test batch from Google. These 300 exam papers were drawn from three sub-categories, namely (case 1), distributed system documents (case 2) and documents processing natural languages (case 3).

The results of the statistical analysis of the three datasets are shown in Fig.4 utilising accurate and recall rate measurements. As expected, the highest rating gives the best accuracy, which demonstrates that the system rating is a solid assessment of the relevance of concepts to documents. There are a lot of elements influencing the performance of the approach. The CPU needed for the preceding experiments is about 2-3 hours duration. The memory needs are rather exorbitant and the number of instances and characteristics are overlapping.

5. CONCLUSION

The fundamental purpose of this work has been the development of a methodology, known as ontological clusters, which can use ontological computing when applied to clustering approaches. In this paper, the system examines how dominant knowledge might be leveraged before the clustering process in the realm of ontology. In addition, these systems do not try to interpret the conceptual meaning of words, which is vital for many textual data applications. The system concentrated on the application of a semantic problem in this work. The evaluation and comparison of concepts is a vital part of the clustering process. This has been done by studying ontological clustering. The usage of ontologies has also been used as more appropriate for the aim of clustering. The system observed that the gradual completeness of ontology is influenced by the clustering findings. Three test cases were evaluated for the evaluation of the system behaviour. The system also discovered that the more precise the ontology, the more precise the results are. The domain ontology has achieved successful outcomes.

REFERENCES

[1] Aqeel-Ur Rehman and Zubair A. Shaikh, "ONTAgri: Scalable Service Oriented Agriculture Ontology for Precision Farming", *Proceedings of International Conference on Agricultural and Biosystems Engineering*, pp. 1-2, 2011.

[2] Boris Lauser, Margherita Sini, Anita Liang, Johannes Keizer and Stephen Katz, "From Agrovoc to the Agricultural Ontology Service/Concept Server", *Proceedings of International Conference on Food and Agriculture Organization of the United Nations*, pp. 1-10, 2006.

[3] Chris Manning and Hinrich Schutze, "*Foundations of Statistical Natural Language Processing*", MIT press, 1999.

[4] Howard Beck, Kelly Morgan, Yunchul Jung, Sabine Grunwald, Ho-Young Kwon and Jin Wu, "Ontology-based Simulation in Agricultural Systems Modeling", *Agricultural Systems*, Vol. 103, No. 7, pp. 463-477, 2010.

[5] Ling Cao and Lin He, "Domain Ontology-based Construction of Agriculture Literature Retrieval System", *Proceeding of International Conference on Wireless*

Communications, Networking and Mobile Computing, pp. 1-3, 2008.

[6] Yuehua Yang, Junping Du and Meiyu Liang, "Study on Food Safety Semantic Retrieval System based on Domain Ontology", *Proceedings of IEEE International Conference on Cloud Computing and Intelligence Systems*, pp. 40-44, 2011.

[7] S. Ningombam, S.P. Meitei and Bipul Syam Purkayastha, "Building Manipuri-English Machine Readable Dictionary by Implementing Ontology", *International Journal of Engineering Science and Technology*, Vol. 3, No. 10, pp. 7682-7689, 2011.

[8] Marwa Hendez and Hadhemi Achour, "Keywords Extraction for Automatic Indexing of E-Learning Resources", *Proceedings of World Symposium on Computer Applications and Research*, pp. 1-5, 2014.

[9] A. Kanaka Durga and A. Govardhan, "Ontology Based Text Categorization-Telugu Document", *International Journal of Scientific and Engineering Research*, Vol. 12, No. 9, pp. 1-4, 2011.

[10] Nigel Collier, Ai Kawazoe, Lihua Jin, Mika Shigematsu, Dinh Dien, Roberto A. Barrero, Koichi Takeuchi and Asanee Kawtrakul, "A Multilingual Ontology for Infectious Disease Surveillance: Rationale, Design and Challenges", *Language Resources and Evaluation*, Vol. 40, pp. 405-413, 2006.

[11] M.R. Priyadarshini, "An Ontology Framework for Context Based Multilingual Document Retrieval", *International Journal of Computer Science and Engineering Technology*, Vol. 5, No. 3, pp. 178-181, 2014.

[12] P. Cimiano, E. Montiel-Ponsoda, P. Buitelaar, M. Espinoza and A. Gomez-Perez, "A Note on Ontology Localization", *Applied Ontology*, Vol. 5, No. 2, pp. 127-137, 2010.

[13] Bo Fu, Rob Brennan and Declan O'Sullivan, "Multilingual Ontology Mapping: Challenges and a Proposed Framework", *Proceedings of Workshop on Matching and Meaning*, pp. 33-35, 2009.

[14] G. Falquet, C. Metral, J. Teller and C. Tweed, "*Ontologies in Urban Development Projects*", 1st Edition, Springer, 2011.

[15] Haytham Al-Feel, Ralph Schafermeier and Adrian Paschke, "An Inter-lingual Reference Approach for Multi-Lingual Ontology", *International Journal of Computer Science*, Vol. 10, No. 2, pp. 497-503, 2013.

[16] C.T. Dos Santos, P. Quaresma and R. Vieira, "An API for Multilingual Ontology Matching", *Proceedings of 7th International Conference on Language Resources and Evaluation*, pp. 3830-3835, 2010.