Ontology based Information Retrieval for Semi Structure Data using Bagging

N.Vanjulavalli,
Research Scholar
Department of Computer Science and Applications
PMU, Thanjavur

A.Kovalan, PhD.
Assistant Professor (S.S),
Department of Computer Science and Applications
PMU, Thanjavur

ABSTRACT

Ontologies are concept specifications and relations that have a major part in semantic web applications through provision of shared knowledge about real world objects ensuring reusability/interoperability among varied modules. So a semantic application should first have an ontology quality related query. Information retrieval (IR) is obtaining information resources relevant to an information need from various information resources. IR has changed over time with expansion of the internet and the arrival of modern graphical user interfaces/ mass storage devices. The aims are using ontologies knowledge to match object with queries on a semantic basis. Ontologies use has many challenges focussing on application of machine learning techniques on features extracted from ontologies concepts and Natural Language Processing. This paper focuses on classifying universities web pages through use of features extracted from an ontology based semantic interpretation.

Keywords

Bagging, Information retrieval (IR), Ontology, World Wide Web.

1. INTRODUCTION

Logic based IR provides a platform to reason about information resources' content meaning in retrieval, i.e. about that meaning's relevance for the user's information need. Thus the user finds resources relevant to his query even without any syntactical similarities. It is clear that retrieval quality depends on domain knowledge's quantity and quality available to the reasoning process [1]. In fact, a logical system retrieves a document on cars for a query on vehicle, only if there is a formal statement that a car is a type of vehicle. Hence, to enable all semantic relevant resources retrieval for a query, domain knowledge has to be acquired and described through a domain theory. Also to resolve "prediction problems," domain theory should be shared, i.e. a common agreement existing about used vocabulary. As ontologies represent explicit domain conceptual specifications, they suit extension of logic-based IR systems as detailed above.

The term ontology is from philosophy, where it is a systematic account of existence. For computer science, what "exists" is that which is represented. Thus, the following definition is adopted [2] in computer science: Ontology is a formal, explicit specification of a shared conceptualisation of a domain of interest.Conceptualisation is an abstract, simplified view of the world that can be represented for a purpose [3]. Ontology plans to overcome issues of implicit and hidden knowledge by ensuring that domain conceptualisation is explicit. It makes assumptions about a specific concept's meaning and can also be an explication of the context for which a concept is used.Moreover, everything (any knowledge-based system/ knowledge-level agent) is liable to conceptualisation either explicitly or implicitly. Hence, it is shared conceptualisation as there is consensus of terms.

Next, ontology aims not to model the whole world, but a part of it - a so called domain which is a specific subject area/area of knowledge like medicine, tool manufacturing, real estate, automobile repair, financial management, etc. Hence, it is important to know what ontology is for to define a domain.Further, ontology establishes a conceptually concise basis for communicating knowledge for varied purposes. Ontology has to be a formal description of the meaning of concepts and relationships between them to achieve this. Hence, formal specification means ontology is specified by a formal language, e.g. firstorder logic.

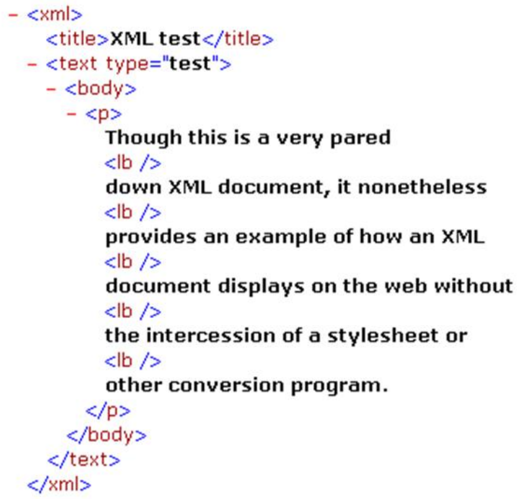
Ontologies are concept specifications and relations among them which play a central role in semantic web applications by providing shared knowledge about real world objects promoting reusability and interoperability among different modules. Thereforethe ontology quality should be the first concern in any semantic application.Ontology-Based Information Extraction (OBIE) is an emerging information extraction sub field. Here, ontologies used by information extraction process with output being generally presented through ontology. Ontology is a formal and explicit specification of a shared conceptualization [4, 5] and they are usually specified for particular domains. As information extraction concerns information retrieval information for a specific domain, specifying that domain's concepts through ontology helps the process [6]. For example, a geopolitical ontology defining concepts like country, province and city guide the information extraction system described earlier. This being the idea on which ontology-based information extraction is based.

In Ontology representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). Representational primitives definition includes information about their meaning and constraints on logically consistent applications. Domain Ontology: models a specific domain representing particular meanings of terms applicable to that domain (CHEMICALS, Gene Ontology).

RDF (Resource Description Framework) converts semantic information into machine accessible information. It is standard for web data interchange with features facilitating data merging even when underlying schemas differ. It specifically supports schemas evolution over time. RDF extends web's linking structure go use URIs to name relationship between things and also the two link ends. This simple model allows structured/semi-structured data to be mixed, exposed, and shared across various applications.

Extensible Mark-up Language (XML) is an emerging standard to represent data on the internet. Sophisticated query engines allowing users to tap data in XML documents is important to exploit XML fully. Though new semi structured data models and query languages were proposed for this purpose, this paper explores ontology use to match object and queries semantically [7].

HTML tags in web pages describe how to display data items while XML tags describe data. The distinction's importance cannot be underestimated. As XML data is self-describing, programs can interpret data meaning that a program receiving an XML document (Figure 1) can interpret it variedly, filter it based on content, restructure it to suit applications etc. This paper showcases features extracted from XML documents based ontologies concepts and Natural Language Processing. The 4 Universities Dataset includes internet pages from computer science departments of major universities which evaluate the proposed method. Feature extraction is compared with features extracted using IDF.



```
- <xml>
    <title>XML test</title>
  - <text type="test">
    - <body>
      - <p>
          Though this is a very pared
          <lb />
          down XML document, it nonetheless
          <lb />
          provides an example of how an XML
          <lb />
          document displays on the web without
          <lb />
          the intercession of a stylesheet or
          <lb />
          other conversion program.
        </p>
      </body>
    </text>
  </xml>
```

Figure 1: Sample XML Document

2. RELATED WORKS

Koopman [8] presented a novel approach to search electronic medical records based on concept matching instead of keyword matching that intends to overcome specific challenges identified when searching medical records. Queries/documents are transformed from term-based originals into medical concepts as defined by SNOMED-CT ontology. A real-world medical records collection evaluation reveals that the new concept-based approach outperforms keyword baseline by 30% in Mean Average Precision. Concept-based approach provides a framework for increased inference based search systems development to deal with medical data.

Q2Semantic [9] bridged the keyword queries and formal queries chasm. The authors had to deal with three problems to achieve their goal including term matching, ranking and scalability. The first is tackled by enriching user queries with Wikipedia terms so that query terms easily match ontology entities. For overcoming the second problem, a ranking mechanism considering factors like query length, relevance and ontology elements importance was implemented. Finally, for scalability issues a clustered graph structure to represent RDF graphs summaries was implemented. But costly graph construction and run-time traversal procedures are still considered necessary.

Paralic [10] presented a new, ontology-based approach to information retrieval (IR) modelled on a domain knowledge representation schema as ontology. Inner system registered resources are linked to concepts from this ontology. Thus, resources may be retrieved based on associations and partial or exact term matching as the vector model user presumes. To evaluate retrieval mechanism quality, retrieval efficiency measuring experiments were performed with well-known CysticFibrosis collection of medical scientific papers. The ontology-based retrieval mechanism was compared to traditional full text search based on vector IR model as well as with the Latent Semantic Indexing method.

Egozi [11] introduced a concept-based retrieval approach based on Explicit Semantic Analysis (ESA), which augments keyword-based text representation with concept-based features, extracted from human knowledge repositories like Wikipedia. The proposed approach automatically generates new text features. It was found that high-quality feature selection was crucial to ensure focussed retrieval. But lack of labelled data, traditional feature selection methods cannot be used and so new methods that use self-generated labelled training data are suggested. The resulting system is evaluated on several TREC datasets and reveals superior performance over earlier state-of-the-art results.

Reymonet [12] described an automotive diagnosis, real-world semantic information retrieval tool. Troubleshooting documents are popular within car work-shops/manufacturers as a method to capitalize on knowledge and also to access repair information. But availability of complex vehicle architectures has led to troubleshooting bases growing so that locating relevant information is harder. Based on a limited knowledge model of automotive diagnosis, our software aims to relieve car mechanics from storing and semantically searching through huge breakdown cases sets.

3. METHODOLOGY

3.1The 4 Universities Dataset

The 4 Universities Dataset includes WWW-pages from computer science departments of major universities collected in January 1997 by the CMU text learning group's World Wide Knowledge Base (Web->Kb) project [13]. The 8,282 pages were manually classified into the following categories: 1) student, 2) faculty, 3) staff, 4) department, 5) course, 6) project and 7) other.

The class other includes pages not deemed the ``main page" representing an instance of earlier six classes. The data set has pages from the four universities for each class: Cornell, Texas, Washington, Wisconsin and 4,120 miscellaneous pages from other universities. The files are collated into a directory structure, one directory for every class. Each of the directories includes 5 subdirectories, one for each of the 4 universities and one for miscellaneous pages. The directories in turn contain Web-pages.

3.2Feature Extraction

Features are extracted using stemming, stop words, finding Inverse Document Frequency (IDF). Inverse Document Frequency (IDF) is measures a word's importance and is defined as the logarithm of the ratio of documents in a collection to the number of documents containing given words [14]. This means rare words possess high IDF and common words low IDF. IDF measures a word's ability to discriminate between documents and is used in many heuristic measures in information retrieval [15]. Document and query are represented as vectors in a high dimensional space corresponding to vector space model's keywords. Similarity measures calculate similarity values between keywords and document and ranking is based on them. The first step is forming a stop word list and stemming words.

The stop word list has non-significant words removed from a document/request before commencing indexing. Stop word list is for words serving no retrieval purpose but used frequently to compose documents,. Such lists are developed so that every match query and document query is based on indexing terms. So document retrieval containing words like "be", "your" and "the" in corresponding request is not a proper search strategy. Non-significant words represent noise, and can damage retrieval as they fail to discriminate between relevant and non-relevant documents, resulting in stop word list listing many pronouns, articles, prepositions and conjunctions. Words like the, a, of, for, with etc., are stop words.

Stemming removes inflectional and derivational suffixes to conflate word variants into the same stem/ root enhancing retrieval effectiveness, assuming that words with similar stem refers to the same idea/concept and hence should be indexed under the same form. To define a stemming algorithm, the first approach removes inflectional suffixes or, for English, this conflates singular/plural word forms and removes past participle ending «-ed» and gerund or present participle ending «-ing». More sophisticated schemes for English corpora were proposed for deviational suffixes (e.g., «-ize», «-ably», «-ship») removal.

Let frequency be denoted by $freq(x,a)$, as it expresses number of occurrences of term a in document x. The term-frequency matrix $TF(x,a)$ measures term a association regarding given document x. $TF(x,a)$ is assigned zero when document does not contain the term, and a number otherwise. The number can be set as $TF(x,a)$= 1 when term a occurs in document x or uses relative term frequency which the frequency versus total occurrences of all document terms. Another measure is **inverse document frequency (IDF)**, representing a scaling factor. If term a occurs frequently in documents, its importance is scaled down due to lowered discriminative power. The $IDF(a)$ is defined as follows: $IDF(a) = \log \frac{1+|x|}{x_a}$

x_a is the set of documents containing term a .Similar documents have same relative term frequencies. Similarity is measured among a document set/between a document and query. Cosine measure locates document similarity [16]; cosine measure is got by:

$$sim(v_1, v_2) = \frac{v_1.v_2}{|v_1|\ |v_2|}$$

Where v_1 and v_2 are two document vectors, $v_1.v_2$ defined as $\sum_{i=1}^{a} v_{1i}v_{2i}$ and $|v_1| = \sqrt{v_1.v_1}$.

Features are extracted based on the ontology in the proposed feature extraction. As ontology specifies a domain conceptualization regarding concepts, attributes, and relations [14], a concept based tree structure (simple hierarchy) is built on a generalisation/specialisation relationship. Difficulties in browsing knowledge bases lead to identifying two requirements: To have a global ontology vision: i.e., to identify the structure (conceptual architecture) of the knowledge base as a whole. To browse knowledge base according to ontology structure: i.e., to access information by exploring intra conceptual hierarchical links. These requirements express problems facing users when browsing in a vast information space. Data mapping provides interesting solutions to this difficulty [17]. Hence, this approach is applicable to semantically annotated knowledge bases resulting in concepts tree structure.

The concepts provided model entities of interest in the domain, organized into a taxonomy tree with each node representing a concept and every concept a specialization of its parent. Figure 2 shows sample taxonomy for Computer Science (CS) department domain (simplifications of real ones) [18].

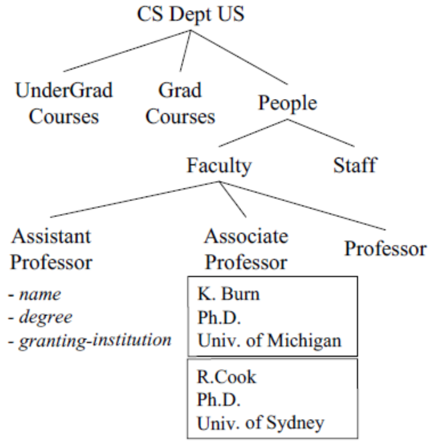


Figure 2: Sample Ontology Tree

3.3 Bagging

Ensemble learning methods goal is constructing a collection (an ensemble) of individual classifiers which are diverse but accurate. When this is possible, then highly accurate classification decisions are also possible by voting the ensemble's individual classifiers decisions. Authors have demonstrated much performance improvement through ensemble methods [19-21].

Two techniques to construct ensembles are bootstrap aggregation [19] and the Adaboost family of algorithms [22]. Both methods use a base learning algorithm and invoke it many times with different training sets. A training set is constructed by forming a bootstrap replica of original training set in bagging. i.e.; given a training set S of m examples, a new training set S' is constructed by drawing m examples uniformly (with replacement) from S.

The Bagging algorithm (Bootstrap aggregating) by Breiman (1996) votes classifiers generated by different bootstrap samples. A Bootstrap sample ensures uniform generation by sampling m instances from training set with replacement. T bootstrap samples B1,….BT are generated and classifier Ci is built from each bootstrap sample Bi. A final classifier C* is built from C1,….CT whose output is most predicted class by sub-classifiers, with arbitrarily broken ties. Bagging algorithm is as follows [20]:

Inputs: training set S, Inducer I, and Number of bootstrap samples T

```
for i= 1 to T {
S' = bootstrap sample from (sample with replacement)
Ci= I(S')
}
```

$$C^{*}\left(x\right)=\arg\max_{y\in Y}\sum_{i:C_i\left(x\right)=y}1$$

Output: classifier C*

In this study, the Bagging is done with REPtree, BFtree, J48, and CART.

4. RESULTS AND DISCUSSION

The proposed ontology based feature extraction for web page classification is evaluated using the 4 Universities Dataset and compared with IDF feature extraction method. The system's performance is evaluated in absolute terms.Classification accuracy, Recall and precision are measured for both proposed semantic and keyword techniques. The accuracy, precision, recall and f measure are computed as follows:

Accuracy (%) = (TN + TP) / (TN + FN + FP + TP)

$$precision=\frac{TP}{TP+FN}$$

$$recall=\frac{TP}{TP+FP}$$

$$f\ Measure=\frac{2*recall*precision}{recall+precision}$$

Where TN (True Negative) = Number of correct predictions that an instance is invalid

FP (False Positive) = Number of incorrect predictions that an instance is valid

FN (False Negative) = Number of incorrect predictions that an instance is invalid

TP (True Positive) = Number of correct predictions that an instance is valid

Keywords and ontology based features are classified using bagging with various decision trees (REPtree, BFtree, J48, and CART). The experimental results obtained are detail in the following tables and figures. Table 1 and Figure 2 details the classification accuracy and root mean squared error obtained for IDF and proposed feature extraction.

Table 1: Classification Accuracy and Root Mean Squared Error

Method Used	Classification Accuracy %	RMS E
Bagging-REPtree-IDF	0.71	0.32
Bagging-BFtree-IDF	0.74	0.31
Bagging-J48-IDF	0.73	0.31
Bagging-CART-IDF	0.77	0.31
Bagging-RREPtree-Proposed feature extraction	0.79	0.28
Bagging-BFtree-Proposed feature extraction	0.83	0.26
Bagging-J48-Proposed feature extraction	0.84	0.24
Bagging-CART-Proposed feature extraction	0.83	0.26

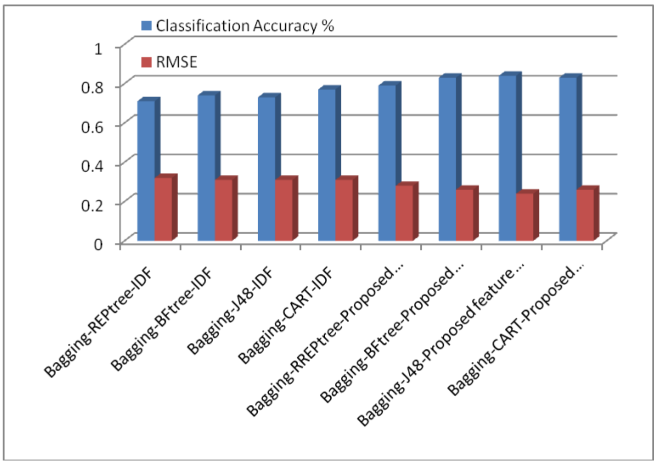


Figure 3: Classification Accuracy and Root Mean Squared Error

It is observed from Figure 3, that the proposed feature extraction performs better than the traditional IDF features. It is also seen that the Root Mean Squared Error is lower for the proposed method. The precision, recall and f measure for the different methods is shown in Table 2 and Figure 4 and 5 shows the precision, recall and f measure respectively.

Table 2: Precision, Recall and F Measure

Method Used	Precision	Recall	f Measure
Bagging-REPtree-IDF	0.715	0.71	0.709
Bagging-BFtree-IDF	0.747	0.74	0.741
Bagging-J48-IDF	0.725	0.73	0.725
Bagging-CART-IDF	0.777	0.77	0.772
Bagging-RREPtree-Proposed feature extraction	0.793	0.79	0.791
Bagging-BFtree-Proposed feature extraction	0.839	0.83	0.833
Bagging-J48-Proposed feature extraction	0.84	0.84	0.84
Bagging-CART-Proposed feature extraction	0.839	0.83	0.833

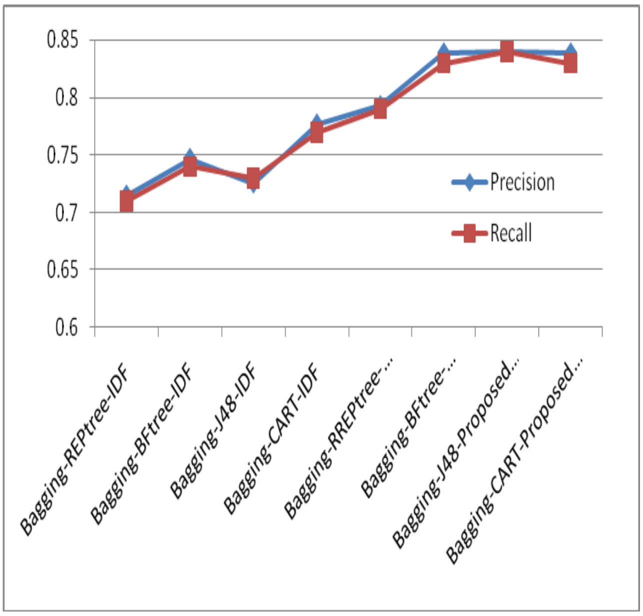


Figure 4: Precision and Recall

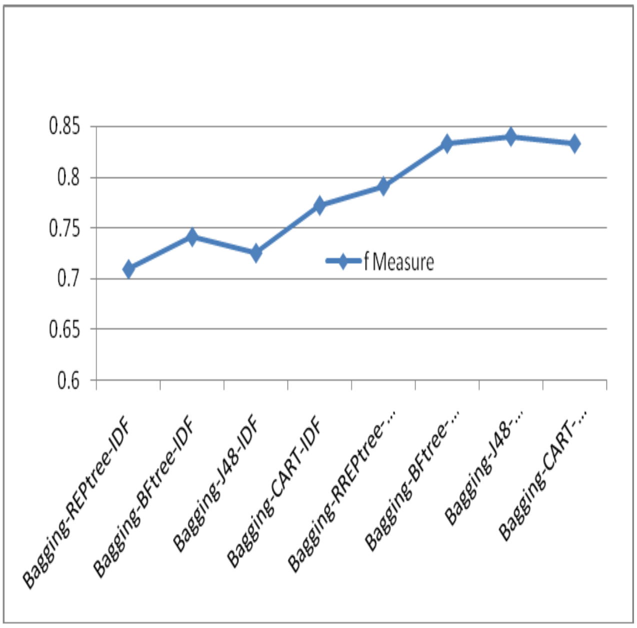


Figure 5: F Measure

The F Measure score is the harmonic mean of recall and precision, a single measure combining recall and precision ensuring that an F Measure score has values within the interval [0, 1]. The F Measure is 0 when relevant documents are not retrieved and 1 when retrieved documents are relevant. Further, harmonic mean F Measure has a high value only when precision and recall are high. Hence, determination of the maximum value for F Measure is an attempt to find a best possible compromise between recall and precision. It is observed from Figure 5 that the combination of Bagging with J48 and the proposed ontology feature extraction achieves the highest F Measure score.

5. CONCLUSION

Information retrieval (IR) is an old research area in information science whose goal is to search and retrieve relevant documents to the user's information needs. Hence, a good IR system should retrieve only documents satisfying user needs and not unnecessary data.Ontology-Based Information Extraction (OBIE) is an emerging information extraction sub field. Here, ontologies used by information extraction process with output being generally presented through ontology.Keywords and ontology based features are classified using bagging with various decision trees (REPtree, BFtree, J48, and CART). The experimental results show that proposed feature extraction improves the precision and recall satisfactorily.

6. REFERENCES

[1] Stojanovic, N., Studer, R., &Stojanovic, L. (2004, September). An approach for step-by-step query refinement in the ontology-based information retrieval. In Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence (pp. 36-43). IEEE Computer Society.

[2] T. R. Gruber, Towards principles for the design of ontologies used for knowledge sharing, in International Journal of Human-Computer Studies, Volume 43, Number 5-6, pp. 907-928, 1995.

[3] N. Guarino, P. Giaretta, Ontologies and Knowledge Bases: Towards a Terminological Clarification, In N. Mars (Eds.): Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, IOS Press, pp. 25-32, 1995.

[4] T. R. Gruber, A translation approach to portable ontology specifications, Knowledge Acquisition 5(2) (1993) 199-220.

[5] R. Studer, V.R. Benjamins and D. Fensel, Knowledge Engineering: Principles and methods, Data Knowledge Engineering 25(1) (1998) 161-197.

[6] Wimalasuriya, D. C., & Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science, 36(3), 306-323.

[7] He, J. S. K. T. G., &Naughton, C. Z. D. D. J. (2008). Relational databases for querying XML documents: Limitations and opportunities. 20.453J / 2.771J / HST.958J Biomedical Information Technology Fall 2008.

[8] Koopman, B., Bruza, P., Sitbon, L., &Lawley, M. (2012). Towards semantic search and inference in electronic medical records: an approach using concept-based information retrieval. Australasian Medical Journal.

[9] Haofen Wang, Kang Zhang, Qiaoling Liu, Thanh Tran, and Yong Yu. Q2semantic: A lightweight keyword interface to semantic search. In ESWC, pages 584–598, 2008.

[10] Paralic, J., &Kostial, I. (2003). Ontology-based information retrieval. In Proceedings of the 14th International Conference on Information and Intelligent systems (IIS 2003), Varazdin, Croatia (pp. 23-28).

[11] Egozi, O., Markovitch, S., &Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. ACM Transactions on Information Systems (TOIS), 29(2), 8.

[12] Reymonet, A., Thomas, J., &Aussenac-Gilles, N. (2009, June). Ontology Based Information Retrieval: an application to automotive diagnosis. In International Workshop on Principles of Diagnosis (DX 2009) (pp. 9-14). Linköping University, Institute of Technology.

[13] Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labelled and unlabeled documents using EM. Machine learning, 39(2), 103-134.

[14] Jones, K. S. (1973). Index term weighting. Information storage and retrieval, 9(11), 619-633.

[15] Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval.

[16] Steinbach, M., Karypis, G., & Kumar, V. (2000, August). A comparison of document clustering techniques. In KDD workshop on text mining (Vol. 400, pp. 525-526).

[17] J. Bertin, 1977. La graphiqueet le traitementgraphique de l'information, Flammarion, Paris.

[18] Doan, A., Madhavan, J., Domingos, P., & Halevy, A. (2004). Ontology matching: A machine learning approach. Handbook on Ontologies in Information Systems, 397-416.

[19] Breiman, L. (1996b). Bias, variance, and arcing classifiers. Tech. rep. 460, Department of Statistics, University of California, Berkeley, CA.

[20] Bauer, E., &Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning, 36 (1/2), 105{139.

[21] Maclin, R., &Opitz, D. (1997). An empirical evaluation of bagging and boosting. In Proceedings of the Fourteenth National Conference on Artificial Intelligence, pp. 546{551 Cambridge, MA. AAAI Press/MIT Press.

[22] Freund, Y., &Schapire, R. E. (1996). Experiments with a new boosting algorithm. In Proc. 13th International Conference on Machine Learning, pp. 148{146. Morgan Kaufmann.