Educational Knowledge Management System Using NLP and Ontology Schema

Pratik S Deshmukh
SEO Analyst, MIT College of Engg, Pune
Kothrud, Pune

Rashmi A. Rane
Asst. Professor, MIT College of Engg, Pune
Kothrud, Pune

ABSTRACT

The three concepts of information science are data, information and knowledge. The structure of one is different from another. The structure of knowledge is more complex than data and information. Knowledge management is complex for traditional information management techniques due to its complex structure and it is difficult to achieve common structure for knowledge captured from heterogeneous sources. Ontology is an upright technology to represent knowledge. Ontology provides homogeneous structure for knowledge acquired from heterogeneous sources. It enables knowledge sharing within and among organizations. Ontology based knowledge management provides a better support for integration of related knowledge sources and searching. The current work proposes an enhanced and clear framework for knowledge management using domain ontology. It addresses major issues of traditional and existing ontology based knowledge management systems. Many Natural Language Processing (NLP) techniques, including stemming, part of-speech tagging, compound recognition, de-compounding, chunking, word sense disambiguation and others, have been used in Information Retrieval (IR). The core IR task we are investigating here is document retrieval. Several other IR tasks use very similar techniques, e.g. document clustering, filtering, new event detection, and link detection, and they can be combined with NLP in a way similar to document retrieval. NLP and IR are very different areas of research, and recent major conferences only have a small number of papers investigating the use of NLP techniques for information retrieval. The moderate success contradicts the intuition that NLP should help IR, which is shared by a large number of researchers. This article reviews the research on combining the two areas and attempts to identify reasons for why NLP has not brought a breakthrough to IR. Natural language processing (NLP) is becoming much more robust and applicable in realistic applications. One area in which NLP has still not been fully exploited is information retrieval (IR).

Keywords
Semantic analysis, ontology, information retrieval.

1. INTRODUCTION

The challenge now is to develop systems capable of simulating human reasoning and enable machines to interpret the information, the ontology by its capacities of representation of the knowledge and the mechanism of reasoning and inference which he offers represents a solution to this need, thus it constitutes the main core of knowledge management systems.

Ontologies are commonly used in the process of information retrieval where the goal is twofold: "understand" the contents of the documents and "understand" the need for the user to be able to put them in relation. Indeed, thanks to the semantics they provide, ontologies may be involved in the reformulation or the extension of user requests or in terms of indexing and annotation of documents and web resources.

This paper’s main objective is the proposal of a knowledge management system based on ontologies. This system has for ambition the capitalization and dissemination of knowledge in a university system. The approach presented here aims to create an organizational memory based on ontologies. The ontology produced will lead to index the documents and thus enable an extension of the classical-based metadata to a search based on semantic criteria. The main content of this ontology result from a manual extraction of the knowledge from a number of documents resulting from the daily work of the university system actors. These documents are validated and then indexed and
classified using ontology. The ontology created will be used to facilitate search and navigation within the field of knowledge capitalized [27].

1.1 Motivation for making use of NLP and ontology model for IR
Imagine we could interact with a university intranet search engine just like with a human person in a natural dialogue. The search engine would automatically extract knowledge from the Web site so that a searcher can be assisted in finding the information required. A student who asks for a particular subject can be directed to the most recent lecture notes or the contact details of the lecturer. An external searcher typing in “PhD NLE” could be assisted by allowing him to explore the space of experts and projects available in the area of natural language engineering. Obviously, this information can change any day and the idea is to have always the most up-to-date facts and relations available to assist a searcher. Currently, we do not have systems which support this type of interaction. However, our aim is to automatically acquire knowledge (a domain model) from the document collection and employ that in an interactive search system [2].

One motivation for a system that guides a user through the search space is the problem of “too many results”. Even queries in document collections of limited size often return a large number of documents, many of them not relevant to the query. Part of the problem is the fact that both on the Web and in intranet search queries tend to be short and short queries always pose ambiguity and uncertainty issues for information retrieval systems. Some form of dialogue based on feedback from the system could be very useful in helping the user find the right results. This combination of NLP and IR we assume is particularly promising and scalable in smaller domains like university intranets or local Web sites.

The knowledge-based theory of the firm suggests that knowledge is the organizational asset that enables sustainable competitive advantage in very dynamic and competitive markets. Knowledge is considered the most important asset for organizations and the effective management of knowledge has become an important issue. KMSs refer to a class of information systems applied to managing organizational knowledge. Knowledge in the context of KMSs consists of experience, know-how and expertise of people (tacit knowledge) as well as different information artifacts, knowledge assets and data stored in documents, reports available within the organization and outside the organization (explicit knowledge).

Knowledge Management Systems are designed to allow users to access and utilize the rich sources of data, information and knowledge stored in different forms, but also to support knowledge creation, knowledge transfer and continuous learning for the knowledge workers. Recently KMSs, unlike databases, have aimed at going beyond the mere administration of electronic information; they now aim at fostering learning processes, knowledge sharing, and collaboration between knowledge workers irrespective of their location, etc. KMSs tend to become complex in order to support the tasks mentioned above as they are not limited to providing easy access to knowledge assets and they address different categories of users with different needs, roles and preferences. Research on user modeling is motivated by two reasons:
1) Differences in individual users needs and
2) Heterogeneity between different groups of users.

Moreover user models and user modeling are the key element for personalized interaction and adaptive feature integration, two very important steps in developing advanced information systems.

2. Literature Survey
In the literature several definitions and visions of knowledge concept are proposed in the area of knowledge management. From there, we adopt a practical and operational vision, which makes a
distinction between data, information and knowledge according to a hierarchical model. In this model, the data are considered as a raw element outside any context. The information is a set of data put in a context. Information is not knowledge, but may become so unless it is understood and assimilated by an individual to perform an action [5].

Note also that some authors made an important distinction between tacit and explicit knowledge. In this perspective, tacit knowledge, as opposed to formal or explicit knowledge, can be integrated into people's heads, in their experience and rooted in action, in the routines in a specific context. Explicit knowledge is knowledge codified and transmitted in a formal and systematic language (documents, information systems, etc.).

Knowledge management covers all the techniques to collect, identify, analyze, and organize, save the knowledge for their sharing and their effective communication between the members of organizations.

Thus, the idea of knowledge management systems is to allow stakeholders to have easy access to knowledge and information sources. This important feature reduces the learning cycles by managing the access to information by disseminating excellence in organizations by sharing best practices, capacity building for learning and innovation, and so on [6].

Ontologies are introduced to cope with many challenges of Knowledge Management Systems (KMS). Ontologies enable KMS to integrate related resources and sharing. Many researchers have proposed and developed ontology based KMSs to handle the knowledge management issues in different fields.


M. Zhou et al., [15] illustrate a general framework for knowledge management from different sources using domain ontology. J. Zhang et al., [12] proposed three multiple ontology based knowledge management system. The detailed survey on the recent ontology based knowledge management approaches, methods and systems concludes that following are the common limitations in the knowledge management.

There is no proper method and clear approach for knowledge management

Many of the knowledge management systems are static

Closed knowledge management

There is no proper approach for uniform representation of heterogeneous data.

2.1 Knowledge Management Systems

KMS refers generally to a system for managing the knowledge within organizations. To build effective KMS technologies, we can say that a KMS could be any of the following:

Document based: It uses technologies that enable the creation, management and sharing of documents such as the Web, distributed databases, document management features, etc.

Ontology based: Knowledge is classified of a set of entities, classes, properties and relations. Moreover, a KMS is supporting knowledge sharing and reuse by covering semantic search methods;

Semantic Web rooted: KMS is an ontology based, they can so be used to explicitly represent semantics of semi-structured and textual information on the web;
Based on AI technologies: Artificial Intelligence techniques are introduced for representing and reasoning about knowledge;
Service based: KMS must deploy knowledge management tools for networks of participants of a project;
Social computing tools are being set up to provide an efficient and natural approach to creation of a KM system. It helps knowledge providers to explicit their implicit knowledge and to formalize knowledge in general [7].

2.2 Conventional IR V.S. Intelligent IR
Traditionally, information retrieval emphasizes document retrieval which is very much dependent on human classification and the use of humanly prepared searching strategies, while intelligent information retrieval emphasizes automatic extraction of useful information and facilitates interaction between the users and facilitates interaction between the user and system by giving the user natural language access tools [8].

2.2.1 Conventional Information Retrieval
In principle, the conventional information retrieval (CIR) systems based on determining the relationships relevant or no relevant between the information need of the users and the information in the documents.. Various extensions to the standard inverted-index technology have been proposed. Those extensions are distance constraints, term weights, synonym specification and term truncation. More recently, because of the explosion of non-bibliographic databases and free style queries requested directly from naïve online searchers, the intelligent information retrieval which is controlled by automatic, machine performed procedures, is needed [9].

2.2.2 Intelligent Information Retrieval
The existence of no bibliographic databases and online information retrieving by computer users pose many problems for information access system, such as, the language for document representation, command language, database selection, problems relating to “user friendliness and ease for use”. Since non-bibliographic databases outnumber their bibliographic counterparts, the need for automatic IR (or Intelligent IR) increases [10].
CIR is rigid, inflexible but fast, portable, relatively inexpensive, relatively easy to learn. IIR is highly expressive, flexible but potentially ambiguous, slow, brittle and expensive [11].

2.2.3 Natural Language in IIR
An information retrieval system without vocabulary control may be referred to as a “natural language” (NL) or “free-text” system. Natural language systems become both more prevalent and more feasible because of the explosion of no bibliographic databases. In addition to free text, IIR might concern with NL queries which are attractive to naïve users who don’t want to learn an artificial query language, which includes Boolean operators, proximity and truncation. However, NL searching offers a number of benefits. That is, it permits the conduct of searches of unlimited specificity [13].

2.3 Ontology or Relational Database
Why would we use an Ontology over a relational database (RDB)? The use of a relational database is to make the following assertion – I understand the data that exists in my domain completely, and that data is relatively static. That is not to say that changes may never happen, but the design of an ERD must remain relatively static for applications to effectively build on top of it. The use of an Ontology model (or models, there is no constraint toward using a single Ontology in an enterprise or industry) is to make the opposite assertion – I do not fully understand the data that exists in my domain, I know that I’ll never understand it completely, and far from being static, the data changes constantly [17].
2.4 Creating an Ontology Model

Semantic models allow users to ask questions about what is happening in a modelled system in a more natural way. As an example, an enterprise might consist of five geographic regions, with each region containing three to five drilling platforms, and each drilling platform monitored by several control systems, each having a different purpose. One of those control systems might monitor the temperature of extracted oil, while another might monitor vibration on a pump. A semantic model will allow a user to ask a question like, "What is the temperature of the oil being extracted on Platform 3?", without having to understand details such as, which specific control system monitors that information or which physical sensor is reporting the oil temperature on that platform [22]. Therefore, semantic models can be used to relate the physical world, as it is known to control systems engineers in this example, to the real world, as it is known to line-of-business leaders and decision makers. In the physical world, a control point such as valve or temperature sensor, is known by its identifier in a particular control system, possibly through a tag name like 14-WW13. This could be one of several thousand identifiers within any given control system, and there could be many similar control systems across an enterprise. To further complicate the problem of information referencing and aggregation, other data points of interest could be managed through databases, files, applications, or component services with each having its own interface method and naming conventions for data accessing [23].

Ontologies are fundamental for Natural Language Processing (NLP) and enhancing search through query expansion at runtime and search index creation on the back-end [24].

3. Proposed work

To provide an interface with a university intranet search engine just like with a human person in a natural dialogue. The search engine would make use of NLP and ontology model to automatically extract knowledge from the database or Web site so that a searcher can be assisted in finding the information required. A student who asks for a particular subject can be directed to the most recent lecture notes or the contact details of the lecturer. A simplified searching technique will be available for the user so that he can make use of natural language to interact with educational knowledge management database without unnecessary search results. Obviously, this information can change any day and the idea is to have always the most up-to-date facts and relations available to assist a searcher by making use of ontology schema of database. Currently, we do not have systems which support this type of interaction. However, my aim is to automatically acquire knowledge (a domain model) from the document collection and employ that in an interactive search system.

3.1 Steps to acquire and process

3.1.1 Mathematical Expression

The Boolean model is based on the manipulation of keywords. On the one side there is a document (D), represented by a combination of keywords, on the other side, a query (R) is represented by a logical expression composed of words and connected by Boolean operators (AND, OR, NOT). The Boolean model uses the exact pairing mode, and it returns only documents corresponding exactly to the query. This model is widely used for bibliographic databases and for web search engines. For a given query first appropriated concepts are retrieved, in this case manually from the user. Then the set of concepts associated with each document is retrieved from database. After that, these two sets are compared using simple metric, which expresses the similarity between a document D and given query Q. where, Qcon is a set of concepts assigned to query Q and Dcon is a set of concepts assigned to document D, and k is a small constant, e.g. 0.1. Resulted number represents ontology-based similarity measure.
Input Data:  
Query related to Educational or University System. e.g. How to get Admission form of ME  
Output:  
Data Retrieval related to Admission form. And applying best filtering We get relevant result for this query.

Problem:  
1. It is not fixed that every time NLP parser will give relevant result.  
2. Result mapping difficult when same word exists.

3.1.2 Use of PROTAGE software for creating ontology schema  
PROTAGE is a free, open source ontology editor and a knowledge acquisition system. Like Eclipse, Protage is a framework for which various other projects suggest plugins. This application is written in Java and heavily uses Swing to create the rather complex user interface.

Firstly sub-domain ontology model of Educational Information (e.g. university management ontology) is constructed. Based on this model, I developed the semantic retrieval system. The relationships of domain concepts can be parsed from the ontology model. Retrieval results can be got from the second step retrieval from the distributed knowledge on the Internet.

3.1.3 Proposed System mapping  
A prototype mapping system that support the process of semi-automatic ontology mapping for the purpose of improving semantic interoperability in heterogeneous systems. The approach is based on the idea of semantic enrichment, i.e. using instance information of the ontology to enrich the original ontologies and calculate similarities between elements in two ontologies. The functional settings for the mapping system are discussed and the evaluation of the prototype implementation of the approach is reported.

The word ontology has been used to describe artifacts with different degrees of structure. These range from simple taxonomies (such as the Yahoo! hierarchy), to metadata schemes (such as the Dublin Core), to logical theories. In our context, the scope and assumption of my work are the following:

(1) An ontology specifies a conceptualization of a domain in terms of concepts, attributes and relations. Concepts are typically organized into a tree structure based on subsumption relationship among concepts. Ad hoc relations further connect concepts and formulate a semantic net structure in the end.

(2) In next stage, the focus will be on finding mappings between concepts and relations. This is because they are the central components of ontologies and matching them successfully would aid in matching the rest of the ontologies.
(3) Ontologies can be expressed in different representational languages. Here, we assume that it is possible to translate between different formats. In practice, a particular representation must be chosen for the input ontologies.

Definition 1 (ontology mapping model)
A 5-tuple \([S, T, F, R(s_i, t_j), A]\) where
- \(S\) is a set composed of logical views (representation) for the elements in source ontology.
- \(T\) is a set composed of logical views (representation) for the elements in target ontology.
- \(F\) is a framework for representing ontology elements and calculating relationships between elements in the two ontologies.
- \(R(s_i, t_j)\) is a ranking function which associate a real number with an element \(s_i \in S\) and an element \(t_j \in T\). Such ranking defines an order among the elements in source ontology with regard to one element \(t_j\) in the target ontology.
- \(A\) is a set composed of mapping assertions. A mapping assertion is a formal description of the mapping result, which supports further description of the exact nature of the derived mappings. It has the following components:
  - A pair of ontology elements,
  - A type of correspondence,
  - A degree of correspondence, and
  - A set of sources of assertion.

3.2 Mapping Technique to be used in future development
Currently we are using The Naive Bayesian classifier which is based on Bayes’ theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

We will propose or develop an algorithm that takes as input semantically enriched elements of ontologies and produces as output suggestions to the user for possible correspondences. The mapper performs a computation of correspondence measures for the pairs of compared ontology elements, based on the similarity of their enriched structures.

- The enhancer utilizes an electronic lexicon to adjust the similarity values that have been computed by the mapper, with the intention of re-ranking the mapping assertions in the result list.
- The presenter determines which recommendations to suggest to the user, based on the partial ordering of correspondence measures and the current configuration profile.
- The exporter translates and exports the mapping results to a desired format so that other follow up applications can import and use the results in a loosely coupled way.
- The configuration profile is a user profile to assign individual variable values for different tuning parameters and a threshold value for exclusion of mappings with low similarity.

3.3 Data independence and Data Flow architecture
NLP Tool Module
Tools included: a sentence splitter, a tokenizer, a part-of-speech tagger, a chunker (used to “find non-recursive syntactic annotations such as noun phrase chunks”), a parser, and a name finder.

Multiplexer Logic
Architected of ontology based Application
3.4 Inputs for the projects
Input: An IR system takes main inputs, the user needs and the information items.
User needs: An information retrieval process begins when a user expresses his information need to the system. In the general case, this information need is conveyed in the form of a search string.
Output: An IR system typically returns one main output, consisting of a minimum relevant list of information items.
The systems that return structured information are commonly characterized as data retrieval systems. While these models do not deal with general information about the subject or topics involved in the sought data, they return very precise answers in response to specific, unambiguous, and formally expressed information needs processes can be identified in NLP tools parse the document, recognize its paragraphs, sentences, and tokens, and provide information about the PoS and the semantic stem of each token.
This information will be used afterwards by the annotation module to discard meaningless tokens such as determinants, prepositions, etc., and to identify lexical structures (tokens or groups of tokens) which might potentially match with ontology entities [30].

3.5 Formal representation of ontology models
Ontology formal representations are explicit description of ontologies formulated in terms of a given ontology language. There have some ontology formal languages such as RDF (Resource Description Framework), RDF(S), OIL, DAML, OWL (DAML+OIL is regarded as a transition), KIF, SHOE, XOL, OCML, Ontolingua, CycL and Loom.
The most widespread ontology language specifically designed for a web-based distributed semantic environment is RDF, which is a recommended standard based on XML by W3C. A simple model is put forward from RDF to express any types of data and the data type consists of marked joint arcs between nodes, which display sources on Web. The arcs display attributes of sources. Therefore, the data model can describe objects and their relationships. RDF data model can be taken as a basic model for any complex model. Therefore, RDF is suitable as a middleware representation language in the application of agricultural sub-domain ontology knowledge [29].

3.6 Ontology models storage
After formal representation, ontology server is needed to provide persistent storage and selective retrieval of ontologies. Unlike common database facilities, ontology server is partially aware of the
semantics part of the data. This can enable ontology server to offer semantic based services. In earlier, semantic data was embedded within HTML or XML pages or simply stored within separate files and served to clients by web servers (e.g. SHOE, OntoBroker). This storage mode is simple. But it is less efficient, it is difficult to adapt to the larger amount of data [26]. At present, the most used model for ontology persistent storage is storing the ontology data into a triple table according the RDF data model. Unlike database facilities in that they are partially aware of the semantics of the data, namely they are able to interpret the standard constructs of ontology languages. This enables ontology servers to offer semantics-based services, such as semantic query. This mode can provide common storage of ontologies, but the query efficient is still low. Efficient persistent ontology storage mode has been a hot research issue in recent years [28].

3.7 Modules Information
Module 1: (GUI Framework and Dataset and Ontology preparation)
In this module the information with respect to the domain as specified above is collected based on the information provided by different universities. This information forms a raw dataset of the collected information. The dataset thus prepared is then processed using an application logic that converts the given raw dataset into a trained ontology.
Module 2: (Query Processing)
This module consists processing on query entered by the user which contains different processes like data pre-processing, lexicon analysis etc.
Module 3: Naive baysan algorithm
Naive baysan algorithm is applied and decision is taken that query is relevant or irrelevant with help of model file that model file is created by using naive baysan algorithm with predefined words. Appropriate words are isolated by using predefined dataset as well as using naive bayesian algorithm and also by using the past history data will be fetched [31].

3.8 Architecture of ontology-based application
One of the important applications of ontology models is to provide semantic query of domain knowledge. A general architecture of ontology based applications is described as figure 5. Where ontology model is key part of applications, after formal representation of ontology models, the model data were stored in the ontology persistent storage facilities. Instead of Semantic server NLP is used to parse, reason or transfer ontology models. The retrieval application is provided to the end users through interface [25].

![Figure 2: A general architecture of ontology based applications](image-url)
4. Conclusion and Future Work

By constructing a framework of knowledge management system based on ontology, this paper expounds the function of each layer, and analyses the implementation of this system from the knowledge organization and expression and knowledge retrieval. Finally, it provides a case which implements the management system and realizes some parts of retrieval modules. This management system establishes a sharable ontology that can be understood both by human and computer, which people can find more relations of different concepts through a better circumstance of knowledge retrieval interface. In addition, the system is also open to some extent, so it can accumulate tacit knowledge constantly and polymerize explicit knowledge efficiently, which can lead to a better management and application of knowledge, to support the innovation for the designers.

In our Research Phase 1, we developed a GUI search engine using NLP for parsing and make use of ontology schema for distributed Educational knowledge management. In phase 2 research we designed a data set of three major universities of India. We created a ontology schema of this dataset to provide searcher exact results to his queries also the comparison between these three universities with respect to limited but important categories. Using Natural language processing (NLP) is becoming much more robust and applicable in realistic applications. This work can obviously only is a first step. There are number of limitations in such a study and I will take the findings as a guideline for future work.

Our current efforts aim at the specification and implementation of the search layer architecture. To put our aims into practice we should first of all develop the domain ontology and study how the content-based similarity between the concepts typed attributes could be assessed. We have been working as well on the design of entirely ontology based structure of the case and the development of our own reasoning methods to operate with it.

REFERENCES


[23] Soma Adindla and Udo Kruschwitz “Combining the Best of Two Worlds: NLP and IR for Intranet Search”, International Conferences on Web Intelligence and Intelligent Agent, University of Essex Wivenhoe Park, Colchester, CO4 3SQ, UK,2011.


[29] Xiaohui Tao, Yuefeng Li, and Ning Zhong,”A Personalized Ontology Model for Web Information Gathering” ,IEEE Transactions on knowledge and data Engineer-ing, Vol.23,No.4,April2011.
