Crossroads
The ACM Student Magazine

Using Perception in Managing Unstructured Documents

by ***Ching Kang Cheng*** and ***Xiaoshan Pan***

Introduction

Over the last ten years, the increased availability of documents in digital form has contributed significantly to the immense volume of knowledge and information available to computer users. The World Wide Web has become the largest digital library available, with more than one billion unique indexable web pages [**12**]. Yet, due to their dynamic nature, fast growth rate, and unstructured format, it is increasingly difficult to identify and retrieve valuable information from these documents. More importantly, the usefulness of an unstructured document is dependent upon the ease and efficiency with which the information is retrieved [**3**]. In this paper, we define an **unstructured document** as a "general" document that is without a specific format e.g., plain text. Whereas, a document divided into sections or paragraph tags is referred to as **semi-structured** e.g., a formatted text document or a web page.

Information management techniques have been developed to analyze large collections of documents, independent of their format. The three most common approaches have focused on information-extraction, information-categorization, and information-retrieval. Although each approach is independent, they can be combined. For example, information-extraction examines the semantics of a document, whereas information-categorization considers the way the document is subdivided. Yet in some cases, techniques employed in information extraction are used to preprocess documents before categorizing them. Information retrieval techniques look into ways to retrieve relevant information from the collection of documents efficiently and effectively. Very often, for optimization purposes, the collection of documents is categorized before applying the information retrieval techniques.

A significant contribution to document management has come from the field of Cognitive Science.

For example, technologies used in Natural Language Processing (NLP) are modeled on human cognition: how humans interpret and understand the semantics in natural language. Together these concepts help form the basis for managing unstructured documents. Here we present a survey of current research and the commercial applications of document management techniques. For greater detail, readers are directed to the referenced articles. It is our intent to give the reader an overview of the available techniques and tools, and their potential usage.

Information Extraction

Natural Language Processing (NLP)

To determine whether or not a document is pertinent to a particular retrieval process, information must be examined in context. This is often accomplished by the technique of NLP. Understanding natural language allows computers to facilitate human problem solving and decision making. Since humans often communicate in a linguistic form, computers that understand natural language can access this information. Natural language computer interfaces allow users to access complex systems intuitively. **syntactic analysis**, **semantic extraction** and **context modeling** are contributing factors in the efficiency and effectiveness of a NLP system. These concepts are explored in greater detail in the following sections.

Syntactic Analysis

Natural language syntax affects the meaning of words and sentences. The meaning of a word varies when syntax is arranged differently. The Link Grammar Parser, developed at Carnegie Mellon University, is based on link grammars, an original theory of English syntax [22]. The parser assigns a valid syntactic structure to a given sentence by connecting a pair of words through a set of labeled links.

The Link Grammar Parser utilizes a dictionary of approximately 60,000 word forms, which comprise a significant variety of syntactic constructions, including many considered rare and/or idiomatic. The parser is robust; it can disregard unrecognizable portions of sentences and assign structures to recognized portions. It can intelligently "guess," from the context, spelling, and probable syntactic categories of unknown words as well. It also considers capitalization, numeric expressions, and various punctuation symbols when making decisions. The Link Grammar Parser can act as the parser in a NLP system.

Semantic Knowledge

Semantic knowledge considers the individual meanings of words and how they integrate in a sentence to gather a collective meaning [1].

Two types of semantic knowledge are essential in a NLP system:

1. **Contextual knowledge**, i.e., how meanings are refined when applied to a specified context.
2. **Lexical knowledge**, or context-independent words (e.g., "children" as the plural form of "child", and

the synonym relationship between "two" and "twice").

WordNet, an electronic lexical database, is one of the most important resources available to researchers. WordNet is used in computational linguistics, text analysis, and other related areas [**9**]. The database WordNet was developed in 1985 by the Cognitive Science Laboratory at Princeton University under the direction of Professor George A. Miller. Its design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets [**16**].

The most basic semantic relationship in WordNet is the synonym. Sets of synonyms, referred to as synsets, form the basic building blocks. Each synset has a unique identifier (ID), a specific definition, and a group of relationships (e.g., inheritance, composition, entailment, etc.) with other synsets.

Ontology and Context Model

An NLP system can only accurately interpret a sentence if it is aware of the context in which the sentence is used. In the following section, the relationship between the user's perspective (**context model**) and NLP can be explained by looking at the characteristics of representable items.

Humans often think in terms of natural language. In Artificial Intelligence (AI), ontologies are developed by humans as models which computers use to perceive the world. An NLP system can only understand text that can be modeled. Direct and indirect mapping relationships exist among vocabularies used by an ontology and vocabularies in a natural language. The quality of the interpretation of free text is strongly dependent on the quality of the model. Coherence, stability, and resistance to inconsistency and ambiguity are desirable ontological model characteristics.

An ontology serves as a representation vocabulary that provides a set of terms with which to describe the facts in some domain. Concepts represented by an ontology can usually be clearly depicted through natural language because the ontology and the natural language function similarly (i.e., describing the world). Most vocabularies used in ontologies are direct subsets of natural languages. For example, a general ontology uses 'thing,' 'entity,' and 'physical;' a specific ontology uses 'dog,' 'car,' and 'tree.'

Depending on the construction of the ontology, the meaning of each word could remain the same as in natural language, or vary completely.

In a natural language, a word may have multiple meanings depending on the applicable context. In a computer system, context may be represented and constrained by an ontology. Vocabularies used in an ontology refer only to the context declared by the ontology. In other words, an ontology provides a context for the vocabulary it contains. Therefore, an ontological model can effectively disambiguate meanings of words from free text sentences.

From the perspective of an NLP system which employs appropriate lexical and contextual knowledge, interpretation of a free text sentence is a process of mapping the sentence from natural language to a context model (**Figure 1**). Different context models may produce varying results simply because words may have different meanings in different contexts.

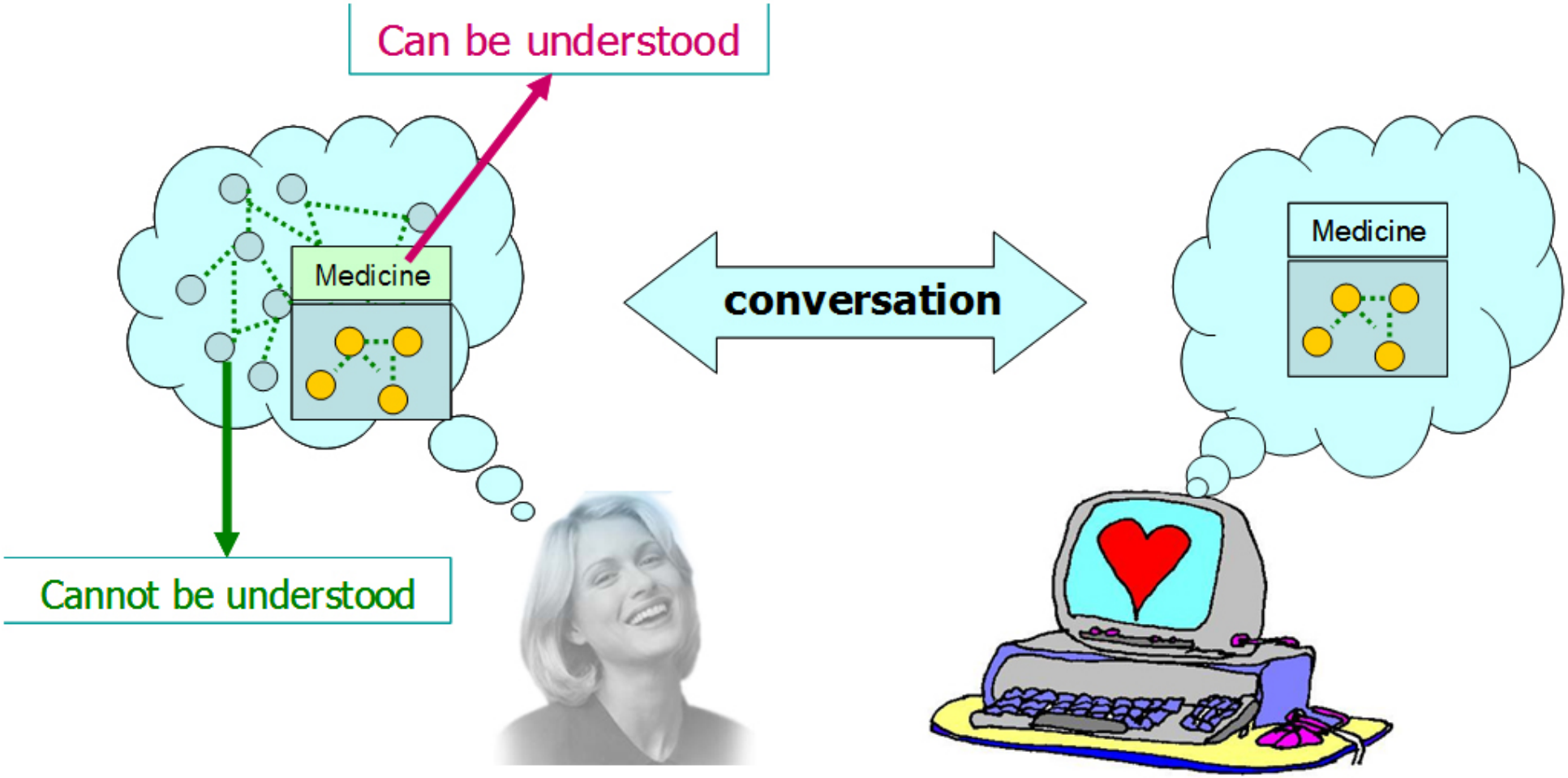

Figure 1: Mapping the sentence from natural language to a context model.

Commercial Application:

The *Semantic Web* is an extension of the current Web. It allows information to be given well-defined meaning, including the semantics of the information. This infrastructure improves the discovery, automation, integration, and sharing of information across various applications [**3**].

In order to support this paradigm, a new kind of markup language is required that allows the definition of common data models or ontologies for a domain, and enables authors to make statements using this ontology. RDF/S and DAML+OIL are markup languages that are currently employed to meet this

need [19]. The introduction of the Semantic Web illustrates the need for machines to interpret the content of a document in context.

Information Categorization

An important aspect in the field of cognitive science is categorization. Humans are naturally good at categorization. When we read a document, we are able to categorize it according to its context. For example, a sports fan can easily classify a web report on the result of a basketball game. On the other hand, to manually categorize information is highly inefficient and often impractical. One deterrent factor is the volume of the information itself. To circumvent this, Information Categorization tools are employed which filter and categorize the collection of documents. These tools are often optimized with the awareness concept [18]; documents are categorized according to the user's perspective. Consequently, human cognitive skills are employed to augment the technologies and ensure better performance.

Information categorization is the process by which documents are classified into different categories. Until the late 1980s, the most common practice was to adopt the knowledge engineering approach. This approach involves manually defining a set of rules with which to categorize a document [21]. In contrast, current technologies in Information Categorization use machine learning (ML). ML is an inductive process that "learns" the characteristics of a category from a set of pre-categorized documents, resulting in an automatic text classifier. An extension of this model is **document clustering**.

Document Clustering

Document clustering is essentially an unsupervised process in which a large collection of text documents are organized into groups of documents that are related, without depending on external knowledge [11]. This has been a challenge. Currently most approaches are grouped into two methods: document partitioning and hierarchical clustering.

The document partitioning methodology consists of two general approaches. First, documents are categorized based on attributes e.g., size, source, topic, and author. Second, the similarity between documents is considered. Documents without similarity are placed into different regions. However, the closer the similarity, the closer the regions are to each other [23].

The hierarchical clustering methods organize the document corpus into a hierarchical tree structure. The clusters in one layer are related with clusters in other layers through association. Document clustering decisions use machine learning algorithms based on **neural networks**.

Neural Networks

Neural networks (NNs), which are based on the theory of cognitive science, simulate biological information processing through parallel, highly-interconnected processing elements (neurons), that cooperate to solve specific problems. Neural networks can be "trained" by example. Generally, NNs are configured for a specific task e.g., data management, via a learning process. The strength of NNs lie

in their ability to derive information from complicated or imprecise data. They can also extract patterns and detect trends too complex for a human or other computer techniques [17].

In addition, NNs can form their own representation of the information they process. Moreover, NNs compute in parallel, allowing special hardware devices to be designed that take advantage of this capability. Consequently, the overall computational time can be shortened. Furthermore, because NNs have a high fault-tolerance, tasks can be performed with incomplete or corrupt data [2].

Major concerns with NNs include scalability, testing, and verification.

Because simulating the parallelism of large problems in sequential machines is difficult, testing and verification of large NNs can be tedious. Since NNs can often function as 'black boxes,' and their internal representation and rules of operation are only partially known, it can be difficult to explain NNs output.

NNs have been used as a learning means for multi-agents in information retrieval [8]. In such cases, each agent learns its environment from users' relevance feedback using a neural network mechanism. The self-organizing map (SOM) [20] is a popular unsupervised neural network model used to automatically structure a document collection.

Commercial Application

Yahoo! (**www.yahoo.com**) groups web sites into categories, creating a hierarchical directory of a subset of the Internet. The hierarchical index created contains more than 150,000 categories (topics) [13]. The popularity and success of Yahoo! demonstrates the strength and potential of information categorization.

Information Retrieval

The ability to retrieve relevant information has been the focus of much research. Three examples are discussed below: **search engines**, **Internet spiders** and **information filtering**. These techniques share the common objective: to assist humans in retrieving the particular bit of information that they need out of the available ocean of information that continues to expand at an astonishing rate. Without these tools, it is almost impossible to depend on human cognition alone for effective and efficient retrieval of relevant information.

Search Engines

A **search engine** optimizes the retrieval process by indexing. Data that is relatively static is preprocessed and stored as a text representation (index) in databases enabling search engines to perform matches more quickly. An elimination technique is often employed to purge frequently occurring words, such as prepositions, which do not contribute to the matching performance but greatly increase the size of the index files.

Another optimization technique uses term-weighting strategies that award higher weights to terms that

are deemed more important during the retrieval of relevant documents. These weights are statistical in nature. Algorithms, therefore, depend on the evaluation of the distribution of terms within individual documents and across the whole document collection [**23**].

Internet Spiders

Internet spiders (a.k.a. crawlers) serve as a vital application in most search engines. The goal of the Internet spider is to gather web pages and at the same time explore the links in each page to propagate the process. Recent years have seen the introduction of client-side web spiders. The shift from running the web spiders on the server-side to the client-side has been popular as more CPU time and memory can be allocated to the search process and greater functionality is possible. More importantly, these tools allow users to have more control and personalization options during the search process. One such feature is the ability to configure a list of web sites to search only relevant sites.

Monitoring and Filtering

More often than not, the contents of web sites are updated frequently. Various tools have been developed to scrutinize web sites for changes and filter out unwanted information. **Push technology** is designed to address such needs. When a user specifies an area of interest a tool will automatically "push" related information to the user. The tool can also be configured to push updates from specified web sites to the user.

Another approach is to employ the use of software agents or intelligent agents. In this case, personalized agents are deployed to track web sites for updates and to filter information according to user needs [**15**]. Machine learning algorithms, such as artificial neural networks, are usually engaged in training the agents to learn the users' preferences.

Commercial Application

The CiteSeer project (**citeseer.nj.nec.com**) finds scientific articles on the Web [**14**]. Information such as an article title, its citations, and their context, is extracted. In addition, full text and autonomous citation indexing are performed. CiteSeer also employs a user profiling system that monitors the interest of users and presents documents as they appear.

Examples of Internet spiders include the **World Wide Web Worm** [**6**], **the Harvest Information Discovery and Access System** [**4**], and the PageRank-based Crawler [**7**]. Focused Crawler [**5**, **12**] is a client-side crawler which locates web pages relevant to a pre-defined set of topics based on example pages provided by the user. Additionally, it has the functionality to analyze the link structures among the web pages collected.

Ewatch (**www.ewatch.com**) monitors information not only from web pages but also from Internet Usenet groups, electronic mailing lists, discussion areas, and bulletin boards to look for changes and alert the user.

Future Research and Commercial Development Trends

Current technologies fail to fully utilize semantic knowledge because they are unable to determine the context of unstructured documents automatically. Today, the semantic of the content can be manually tagged in Extensible Markup Language (XML) with the unstructured document. Such an approach is severely limited as it is not scalable nor efficient and requires users to know the overall structure of a document or its exact name and form in advance.

We envision future research to focus in the area of integrating users' context when retrieving information from unstructured documents. The Semantic Web is one possible approach, in which pages can be given well-defined meaning. Software agents can also assist web users by using this information to search, filter, and prepare information in new ways [**10**]. Besides improving the quality of the search, such an approach allows better integration between machines and people and assists the evolution of human knowledge as a whole [**3**]. In addition, future technologies must have the capability to automatically extract the meaning of the unstructured documents with reference to the context of the users and with minimal human intervention.

Knowledge encompassed in unstructured documents can reach its full potential only if it can be shared and processed by automated tools as well as by people. Furthermore, to ensure scalability, tomorrow's programs must be able to share and process information even when these programs have been designed totally independently.

References

1

Allen, J. *Natural Language Understanding*. Redwood City, California: Benjamin/Cummings Publishing Company, 1995.

2

Becks, A., Sklorz, S., and Jarke, M. A Modular Approach for Exploring the Semantic Structure of Technical Document Collections. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, May 2000.

3

Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. *Scientific American*, May 2001.

4

Bowman, C. and Danzig, P. The Harvest Information Discovery and Access System. In *Proceedings of the Second International World-Wide Web Conference*, October 1994.

5

Chakrabarti, S., van der Berg, M., and Dom, B. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. In *Proceedings of the 8th International World Wide Web Conference*, 1999.

6

Chau, M., Chen, H., Qin, J., Zhou, Y., Qin, Y., Sung, W., and McDonald, D. Novel Search Environments: Comparison of Two Approaches to Building a Vertical Search Tool: A Case Study in the Nanotechnology Domain. In *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*, July 2002.

7

Cho, J., Garcia-Molina, H., and Page, L. Efficient Crawling Through URL Ordering. In *Proceedings of the 7th World Wide Web Conference Brisbane*, Australia, April 1998.

8

Choi, Y, S. and Yoo, Y. S. Multi-agent Learning Approach to WWW Information Retrieval Using Neural Network. In *Proceedings of the 4th International Conference on Intelligent User Interfaces*, December 1998.

9

Fellbaum, C. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press, 1999.

10

Hendler, J., Berners-Lee, T., and Miller, E. Integrating Applications on the Semantic Web. *Journal of the Institute of Electrical Engineers of Japan*, Volume 122(10), pp. 676-680, October 2002.

11

Kim, H. J. and Lee S. G. A Semi-supervised Document Clustering Technique for Information Organization. In *Proceedings of the ninth international conference on Information and knowledge management*, McLean, Virginia, 2000.

12

Kobayashi, M. and Takeda, K. Information Retrieval on the Web. *ACM Computing Surveys (CSUR)*. Volume 32, Issue 2, June 2000.

13

Labrou, Y., Finin, T. Yahoo! as an Ontology: Using Yahoo! Categories to Describe Documents. In *Proceedings of the eighth international conference on Information and knowledge management*, November 1999.

14

Lawrence, S., Bollacker, K., and Giles, C. L. Indexing and Retrieval of Scientific Literature. In *Proceedings of the eighth international conference on Information and knowledge management*, November 1999.

15

Maes, P. Agents that Reduce Work and Information Overload. *Communications of the ACM*, 37(7), 1994, pp. 31-40.

16

Miller, G. Wordnet: An Online Lexical Database. *International J. Lexicography*, Vol. 3, No. 4, 1990, pp. 235-312.

17

N. E. Sondak , V. K. Sondak. Neural Networks and Artificial Intelligence. In *ACM SIGCSE Bulletin, Proceedings of the Twentieth SIGCSE Technical Symposium on Computer Science Education*. Volume 21, Issue 1, February 1989.

18

Nardi, B, A. Awareness Essay: Concepts of Cognition and Consciousness: Four Voices. *ACM SIGDOC Asterisk Journal of Computer Documentation*. Volume 22, Issue 1, February 1998.

19

Patel-Schneider, P., and Sim´on, J. Languages & Authoring for the Semantic Web: The Yin/Yang web: XML syntax and RDF semantics. In *Proceedings of the eleventh international conference on World Wide Web*, May 2002.

20

Rauber, A. and Merkl, D. SOMLib: A Digital Library System Based on Neural Networks. In *Proceedings of the fourth ACM conference on Digital libraries*, August 1999.

21

Sebastiani, F. Machine Learning in Automated Text Categorization. *ACM Computing Surveys (CSUR)*. Volume 34 Issue 1, March 2002.

22

Sleator, D. and Temperley, D. Parsing English with a Link Grammar. *Carnegie Mellon University Computer Science Technical Report CMU-CS-91-196*, 1991.

23

Yang, H. and M, Palaniswami. On the Issue of Neighborhood in Self-organizing Maps. In *Proceedings of the 1992 ACM/SIGAPP Symposium on Applied Computing: Technological Challenges of the 1990's*, April 1992.

Biographies

Ching Kang Cheng (**ckcheng@calpoly.edu**) is a graduate student at California Polytechnic State University, San Luis Obispo, working towards his MS in Computer Science. His research interests include Knowledge Management, Knowledge Representation, and Multi-agents Systems.

Xiaoshan Pan (**xpan@stanford.edu**) is a graduate student pursuing a PhD from the Department of Civil and Environmental Engineering at Stanford University. His research interests include Machine Learning, Natural Language Processing, Complex Adaptive Systems, and Multi-agent Systems.