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 indigital form has contributed significantly to the immense volume ofknowledge and information available to computer users. The WorldWide Web has become the largest digital library available, withmore 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 identifyand retrieve valuable information from these documents. Moreimportantly, the usefulness of an unstructured document isdependent upon the ease and efficiency with which the informationis retrieved [**3**]. In this paper, we define an**unstructured document** as a "general" document that iswithout a specific format e.g., plain text. Whereas, a documentdivided into sections or paragraph tags is referred to as**semi-structured** e.g., a formatted text document or a webpage.

Information management techniques have been developed to analyzelarge collections of documents, independent of their format. Thethree most common approaches have focused oninformation-extraction, information-categorization, andinformation-retrieval. Although each approach is independent, theycan be combined. For example, information-extraction examines thesemantics of a document, whereas information-categorizationconsiders the way the document is subdivided. Yet in some cases,techniques employed in information extraction are used topreprocess documents before categorizing them. Informationretrieval techniques look into ways to retrieve relevantinformation from the collection of documents efficiently andeffectively. Very often, for optimization purposes, the collectionof documents is categorized before applying the informationretrieval techniques.

A significant contribution to document management has come fromthe field of Cognitive Science. For example, technologies used inNatural Language Processing (NLP) are modeled on human cognition: how humans interpret and understand the semantics in naturallanguage. Together these concepts help form the basis for managingunstructured documents. Here we present a survey of currentresearch and the commercial applications of document managementtechniques. For greater detail, readers are directed to thereferenced articles. It is our intent to give the reader anoverview of the available techniques and tools, and their potentialusage.

Information Extraction

Natural Language Processing (NLP)

To determine whether or not a document is pertinent to a particularretrieval process, information must be examined in context. This isoften accomplished by the technique of NLP. Understanding naturallanguage allows computers to facilitate human problem solving anddecision making. Since humans often communicate in a linguisticform, computers that understand natural language can access thisinformation. Natural language computer interfaces allow users toaccess complex systems intuitively. **syntactic analysis**,**semantic extraction** and **context modeling** arecontributing factors in the efficiency and effectiveness of a NLPsystem. These concepts are explored in greater detail in thefollowing 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 structureto a given sentence by connecting a pair of words through a set oflabeled links.

The Link Grammar Parser utilizes a dictionary of approximately60,000 word forms, which comprise a significant variety ofsyntactic constructions, including many considered rare and/oridiomatic. The parser is robust; it can disregard unrecognizableportions of sentences and assign structures to recognized portions.It can intelligently "guess," from the context, spelling, andprobable syntactic categories of unknown words as well. It alsoconsiders capitalization, numeric expressions, and variouspunctuation symbols when making decisions. The Link Grammar Parsercan act as the parser in a NLP system.

Semantic Knowledge

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

Two types of semantic knowledge are essential in a NLPsystem:

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

WordNet, an electronic lexical database, is one of the mostimportant resources available to researchers. WordNet is used incomputational linguistics, text analysis, and other related areas[9].

The database WordNet was developed in 1985 bythe Cognitive Science Laboratory at Princeton University under thedirection of Professor George A. Miller. Its design is inspired bycurrent psycholinguistic theories of human lexical memory. Englishnouns, verbs, and adjectives are organized into synonym sets, eachrepresenting one underlying lexical concept. Different relationslink the synonym sets [16].

The most basic semantic relationship in WordNet is the synonym.Sets of synonyms, referred to as synsets, form the basic buildingblocks. Each synset has a unique identifier (ID), a specificdefinition, 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 isaware of the context in which the sentence is used. In thefollowing section, the relationship between the user's perspective(**context model**) and NLP can be explained by looking at thecharacteristics of representable items.

Humans often think in terms of natural language. In ArtificialIntelligence (AI), ontologies are developed by humans as modelswhich computers use to perceive the world. An NLP system can onlyunderstand text that can be modeled. Direct and indirect mappingrelationships exist among vocabularies used by an ontology andvocabularies in a natural language. The quality of theinterpretation of free text is strongly dependent on the quality ofthe model. Coherence, stability, and resistance to inconsistencyand ambiguity are desirable ontological model characteristics.

An ontology serves as a representation vocabulary that providesa set of terms with which to describe the facts in some domain.Concepts represented by an ontology can usually be clearly depictedthrough natural language because the ontology and the naturallanguage function similarly (i. e., describing the world). Mostvocabularies used in ontologies are direct subsets of naturallanguages. 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 ofeach word could remain the same as in natural language, or varycompletely.

In a natural language, a word may have multiple meaningsdepending on the applicable context. In a computer system, contextmay be represented and constrained by an ontology. Vocabulariesused in an ontology refer only to the context declared by theontology. In other words, an ontology provides a context for thevocabulary it contains. Therefore, an ontological model caneffectively disambiguate meanings of words from free textsentences.

From the perspective of an NLP system which employs appropriatelexical and contextual knowledge, interpretation of a free textsentence is a process of mapping the sentence from natural languageto a context model (**Figure 1**). Differentcontext models may produce varying results simply because words mayhave different meanings in different contexts.

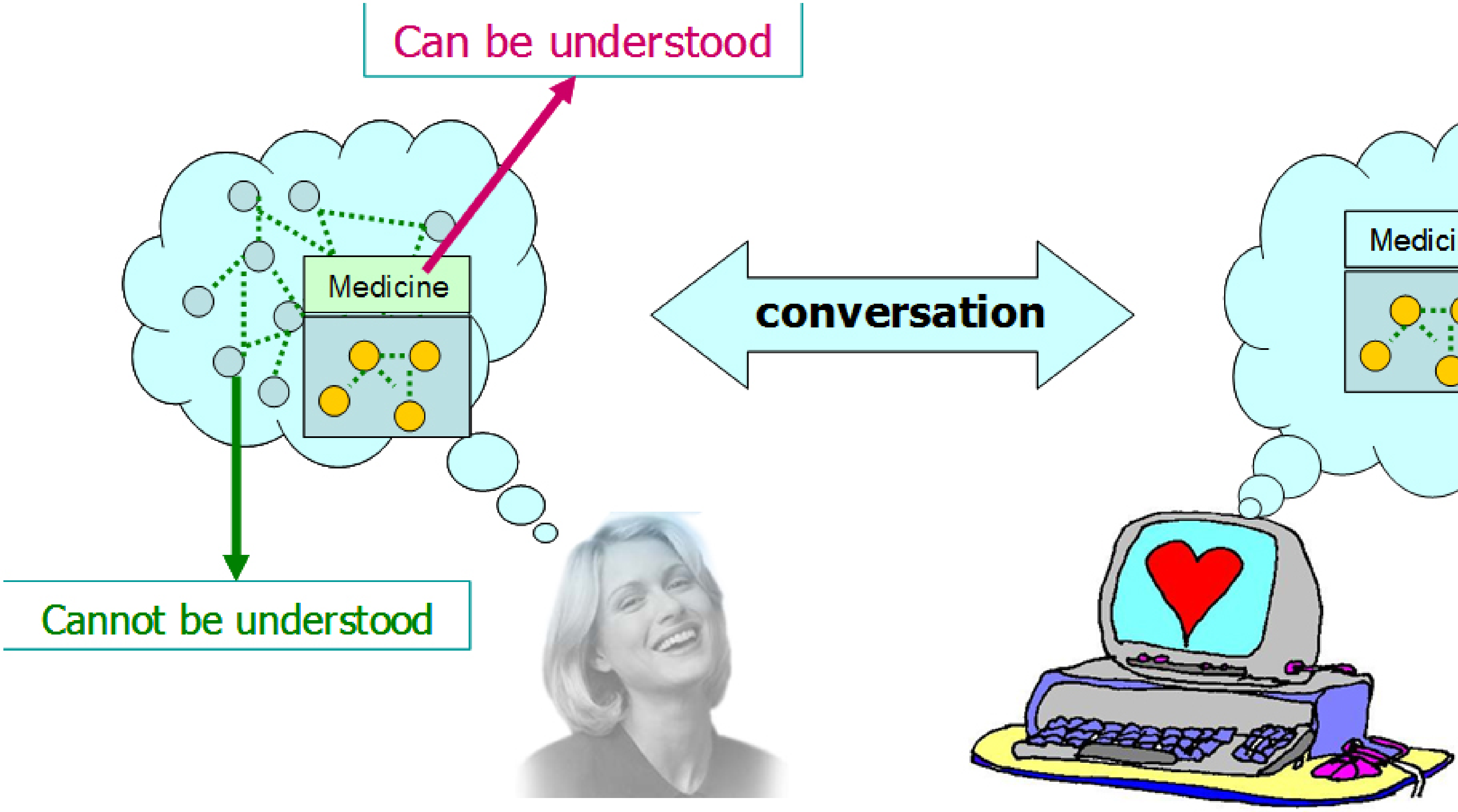

Figure 1: Mapping the sentence from natural language to a context model.

Commercial Application:

The *Semantic Web* is an extension of the current Web. Itallows information to be given well-defined meaning, including thesemantics of the information. This infrastructure improves thediscovery, automation, integration, and sharing of informationacross various applications [3].

In order to support this paradigm, a new kind of markup languageis required that allows the definition of common data models orontologies for a domain, and enables authors to make statementsusing this ontology. RDF/S and DAML+OIL are markup languages thatare currently employed to meet this need [19].The introduction of the Semantic Web illustrates the need formachines to interpret the content of a document in context.

Information Categorization

An important aspect in the field of cognitive science iscategorization. Humans are naturally good at categorization. Whenwe read a document, we are able to categorize it according to itscontext. For example, a sports fan can easily classify a web reporton the result of a basketball game. On the other hand, to manuallycategorize information is highly inefficient and often impractical.One deterrent factor is the volume of the information itself. Tocircumvent this, Information Categorization tools are employedwhich filter and categorize the collection of documents. Thesetools are often optimized with the awareness concept [18]; documents are categorized according to the user'sperspective. Consequently, human cognitive skills are employed toaugment the technologies and ensure better performance.

Information categorization is the process by which documents areclassified into different categories. Until the late 1980s, themost common practice was to adopt the knowledge engineeringapproach. This approach involves manually defining a set of ruleswith which to categorize a document [21]. Incontrast, current technologies in Information Categorization usemachine learning (ML). ML is an inductive process that "learns" thecharacteristics of a category from a set of pre-categorizeddocuments, resulting in an automatic text classifier. An extensionof this model is **document clustering**.

Document Clustering

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

The document partitioning methodology consists of two generalapproaches. First, documents are categorized based on attributese.g., size, source, topic, and author. Second, the similaritybetween documents is considered. Documents without similarity areplaced into different regions. However, the closer the similarity,the closer the regions are to each other [23].

The hierarchical clustering methods organize the document corpusinto a hierarchical tree structure. The clusters in one layer arerelated with clusters in other layers through association. Documentclustering decisions use machine learning algorithms based on**neural networks**.

Neural Networks

Neural networks (NNs), which are based on the theory of cognitivescience, simulate biological information processing throughparallel, highly-interconnected processing elements (neurons), thatcooperate to solve specific problems. Neural networks can be"trained" by example. Generally, NNs are configured for a specifictask e.g., data management, via a learning process. The strength ofNNs lie in their ability to derive information from complicated orimprecise data. They can also extract patterns and detect trendstoo complex for a human or other computer techniques [17].

In addition, NNs can form their own representation of theinformation they process. Moreover, NNs compute in parallel,allowing special hardware devices to be designed that takeadvantage of this capability. Consequently, the overallcomputational time can be shortened. Furthermore, because NNs havea high fault-tolerance, tasks can be performed with incomplete orcorrupt data [2].

Major concerns with NNs include scalability, testing, andverification.

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

NNs have been used as a learning means for multi-agents ininformation retrieval [8]. In such cases, eachagent learns its environment from users' relevancefeedback using a neural network mechanism. The self-organizing map(SOM) [20] is a popular unsupervised neuralnetwork model used to automatically structure a documentcollection.

Commercial Application

Yahoo! (**www.yahoo.com**) groupsweb sites into categories, creating a hierarchical directory of asubset of the Internet. The hierarchical index created containsmore than 150,000 categories (topics) [13]. Thepopularity and success of Yahoo! demonstrates the strength andpotential of information categorization.

Information Retrieval

The ability to retrieve relevant information has been the focus ofmuch research. Three examples are discussed below: **searchengines**, **Internet spiders** and **informationfiltering**. These techniques share the common objective: toassist humans in retrieving the particular bit of information thatthey need out of the available ocean of information that continuesto expand at an astonishing rate. Without these tools, it is almostimpossible to depend on human cognition alone for effective andefficient 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 textrepresentation (index) in databases enabling search engines toperform matches more quickly. An elimination technique is oftenemployed to purge frequently occurring words, such as prepositions,which do not contribute to the matching performance but greatlyincrease the size of the index files.

Another optimization technique uses term-weighting strategiesthat award higher weights to terms that are deemed more importantduring the retrieval of relevant documents. These weights arestatistical in nature. Algorithms, therefore, depend on theevaluation of the distribution of terms within individual documentsand across the whole document collection [23].

Internet Spiders

Internet spiders (a.k.a. crawlers) serve as a vitalapplication in most search engines. The goal of the Internet spideris to gather web pages and at the same time explore the links ineach page to propagate the process. Recent years have seen theintroduction of client-side web spiders. The shift from running theweb spiders on the server-side to the client-side has been popularas more CPU time and memory can be allocated to the search processand greater functionality is possible. More importantly, thesetools allow users to have more control and personalization optionsduring the search process. One such feature is the ability toconfigure a list of web sites to search only relevant sites.

Monitoring and Filtering

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

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

Commercial Application

The CiteSeer project (**citeseer.nj.nec.com**) findsscientific articles on the Web [14]. Informationsuch as an article title, its citations, and their context, isextracted. In addition, full text and autonomous citation indexingare performed. CiteSeer also employs a user profiling system thatmonitors the interest of users and presents documents as theyappear.

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

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

Future Research and Commercial Development Trends

Current technologies fail to fully utilize semantic knowledgebecause they are unable to determine the context of unstructureddocuments automatically. Today, the semantic of the content can bemanually tagged in Extensible Markup Language (XML) with theunstructured document. Such an approach is severely limited as itis not scalable nor efficient and requires users to know theoverall structure of a document or its exact name and form inadvance.

We envision future research to focus in the area of integratingusers' context when retrieving information from unstructureddocuments. The Semantic Web is one possible approach, in whichpages can be given well-defined meaning. Software agents can alsoassist web users by using this information to search, filter, andprepare information in new ways [10]. Besidesimproving the quality of the search, such an approach allows betterintegration between machines and people and assists the evolutionof human knowledge as a whole [3]. In addition,future technologies must have the capability to automaticallyextract the meaning of the unstructured documents with reference tothe context of the users and with minimal human intervention.

Knowledge encompassed in unstructured documents can reach itsfull potential only if it can be shared and processed by automatedtools as well as by people. Furthermore, to ensure scalability, tomorrow's programs must be able to share and process informationeven when these programs have been designed totallyindependently.

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 forExploring the Semantic Structure of Technical Document Collections.In *Proceedings of the Working Conference on Advanced VisualInterfaces*, 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 andAccess System. In *Proceedings of the Second InternationalWorld-Wide Web Conference*, October 1994.

5

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

6

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

7

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

8

Choi, Y, S. and Yoo, Y. S. Multi-agent Learning Approach to WWWInformation Retrieval Using Neural Network. In *Proceedings ofthe 4th International Conference on Intelligent UserInterfaces*, December 1998.

9

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

10

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

11

Kim, H. J. and Lee S. G. A Semi-supervised Document ClusteringTechnique for Information Organization. In *Proceedings of theninth international conference on Information and knowledgemanagement*, McLean, Virginia, 2000.

12

Kobayashi, M. and Takeda, K. Information Retrieval on the Web.*ACM Computing Surveys (CSUR)*. Volume 32, Issue 2, June2000.

13

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

14

Lawrence, S., Bollacker, K., and Giles, C. L. Indexing andRetrieval of Scientific Literature. In *Proceedings of the eighthinternational conference on Information and knowledgemanagement*, 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 ArtificialIntelligence. In *ACM SIGCSE Bulletin, Proceedings of theTwentieth SIGCSE Technical Symposium on Computer ScienceEducation*. Volume 21, Issue 1, February 1989.

18

Nardi, B, A. Awareness Essay: Concepts of Cognition andConsciousness: Four Voices. *ACM SIGDOC Asterisk Journal ofComputer 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 andRDF semantics. In *Proceedings of the eleventh internationalconference on World Wide Web*, May 2002.

20

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

21

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

22

Sleator, D. and Temperley, D. Parsing English with a LinkGrammar. *Carnegie Mellon University Computer Science TechnicalReport CMU-CS-91-196*, 1991.

23

Yang, H. and M, Palaniswami. On the Issue of Neighborhood inSelf-organizing Maps. In *Proceedings of the 1992 ACM/SIGAPPSymposium on Applied Computing: Technological Challenges of the1990's*, April 1992.

Biographies

Ching Kang Cheng (**ckcheng@calpoly.edu**) is agraduate student at California Polytechnic State University, SanLuis Obispo, working towards his MS in Computer Science. Hisresearch interests include Knowledge Management, KnowledgeRepresentation, and Multi-agents Systems.

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