Construction and Visualization of Key Term Hierarchies

Joe Zhou and Troy Tanner
LEXIS-NEXIS, a Division of Reed Elsevier
9555 Springboro Pike
Miamisburg, OH 45342
{joez, tlt}@lexis-nexis.com

Abstract

This paper presents a prototype system for key term manipulation and visualization in a real-world commercial environment. The system consists of two components. A preprocessor generates a set of key terms from a text dataset which represents a specific topic. The generated key terms are organized in a hierarchical structure and fed into a graphic user interface (GUI). The friendly and interactive GUI toolkit allows the user to visualize the key terms in context and explore the content of the original dataset.

1. INTRODUCTION

As the amount of on-line text grows at an exponential rate, developing useful text analysis techniques and tools to access information content from various electronic sources is becoming increasingly important. In this paper we present an applied research prototype system that intends to accomplish two major tasks. First, a set of key terms, ranging from single word terms to four word terms, are automatically generated and organized in a hierarchical structure out of a text dataset which represents a specific topic. Second, a graphic user interface (GUI) is established that provides the domain expert or the user with an interactive environment to visualize the key term hierarchy in the context of the original dataset.

2. SYSTEM DESCRIPTION

The ultimate goal of this prototype system is to offer an automated toolkit which allows the domain expert or the user to visualize and examine key terms in a large information collection. Such a toolkit has proven to be useful in a number of real applications. For example, it has helped us reduce the time and manual effort needed to develop and maintain our on-line document indexing and classification schemes.

The system consists of two components: a preprocessing component for the automatic construction of key terms and the front-end component for user-guided graphic interface.

2.1 Automatic Generation of Key Terms

Automatically identifying meaningful terms from naturally running texts has been an important task for information technologists. It is widely believed that a set of good terms can be used to express the content of the document. By capturing a set of good terms, for example, relevant documents can be searched and retrieved from a large document collection. Though what constitutes a good term still remains to be answered, we know that a good term can be a word stem, a single word, a multiple word term (a phrase), or simply a syntactic unit.

Various existing and workable term extraction tools are either statistically driven, or linguistically oriented, or some hybrid of the two. They all target frequently co-occurring words in running text. The earlier work of Choueka (1988) proposed a pure frequency approach in which only quantitative selection criteria were established and applied. Church and Hanks (1990) introduced a statistical measurement called mutual information for extracting strongly associated or collocated words. Tools like Xtract (Smadja 1993) were based on the work of Church and others, but made a step forward by incorporating various statistical measurements like z-score and variance of distribution, as well as shallow linguistic techniques like part-of-speech tagging and lemmatization of input data and partial parsing of raw output. Exemplary linguistic approaches can be found in the work by Strzalkowsky (1993) where a fast and accurate syntactic parser is the prerequisite for the selection of significant phrasal terms.

Different applications aim at different types of key terms. For the purpose of generating key terms for our prototype system, we have adopted a "learn data from data" approach. The novelty of this approach lies in the automatic comparison of two sample datasets, a topic focused dataset based on a predefined topic and a larger and more general base dataset. The focused dataset is created by the domain expert either through a submission of an on-line search or through a compilation of documents from a specific source. The construction of the corresponding base dataset is performed by pulling documents out of a number of sources, such as news wires, newspapers, magazines and legal databases. The intention is to make the resulted corpora cover a much greater variety of topics or domain subjects than the focused dataset.

To identify interesting word patterns in both samples a set of statistical measures are applied. The identification of single word terms is based on the variation of a t-test. Two-word terms are captured through the computation of mutual information (Church et al. 1991), and an extension of mutual information assists in extracting three-word and four-word terms. Once the significant terms of these four types are identified, a comparison algorithm is applied to differentiate terms across the two samples. If significant changes in the values of certain statistical variables are detected, associated terms are selected from the focused sample and included in the final generated lists. (For a complete description of the algorithm and preliminary experiments, please refer to Zhou and Dapkus 1995.)

2.2 Graphic User Interface (GUI)

We view our prototype system as a means to achieve information visualization. Analogous to scientific visualization that allows scientists to make sense out of intellectually large data collections, information visualization aims at organizing large information spaces so that information technologists can visualize what is out there and how various parts are related to each other (Robertson et al. 1991). The guiding principle for building the GUI component of our prototype system is to automate the manual process of capturing information content out of large document collections.

2.2.1 General Presentation

The design of the GUI component relies on a number of well understood elements which include a suggestive graphic design and a direct manipulation metaphor to achieve an easy-to-learn user interface. The layout of the graphic design is intended to facilitate the quick comprehension of the displayed information. The GUI component is divided into two main areas, one for interacting with key terms structures and one for browsing targeted document collections.

The following descriptions should be viewed together with the appropriate figures of the GUI component. Figure 1, attached at the end of the paper, represents the overall GUI picture. Figures 2 and 3 capture the area where the interaction with the key term structures occurs. Figures 4 and 5 present the area for document browsing and key terms selection. The topic illustrated in the figures is the legal topic "Medical Malpractice".

2.2.2 Term Access Mechanism

The left area of the GUI component (see figures 2 and 3) is devoted to selecting, retrieving and operating on the key terms generated by the preprocessing component of the prototype system. As can be seen, the key terms, ranging from single word terms to four word terms, are organized in a tree structure. The tree is a two dimensional visualization of the term hierarchy. Single word terms are represented as root nodes and multiple word terms can be positioned uniformly below the parent node in the term hierarchy. The goal of the visualization is to present the key term lists in such a way that a high percentage of the hierarchy is visible with minimal scrolling.

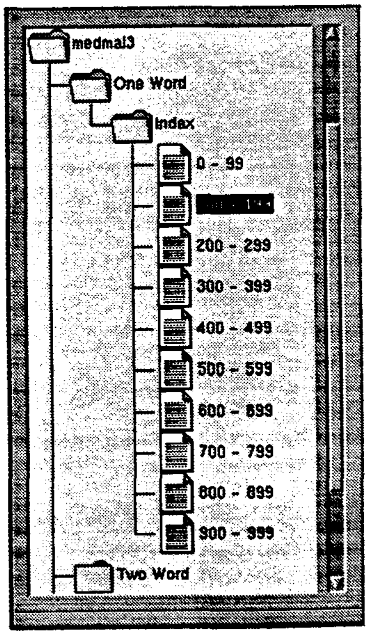


Figure 2

The user interaction is structured around term retrieval and navigation as the top level user interactions. The retrieval of the key terms is treated as an iterative process in which the user may select single world terms from the term hierarchy and navigate to multiple word terms accordingly.

The user begins term navigation by selecting from a list of available topics. In this case, the legal topic "Medical Malpractice" (i.e., medmal3) is selected (see figure 2). Often data structures are organized linearly by some metric. Frequency of key term usage is the metric used to organize and partition the term hierarchy in an ascending numerical order. The partitioning is necessary as it is difficult to accommodate the large ratio of the term hierarchy on the screen. Currently, each partition contains 100 root nodes (or folders), representing single word terms. Once a partition has been selected, the corresponding document collection is loaded into the document browser. The browser provides the user with the ability to quickly navigate through the document collection to locate relevant key terms.

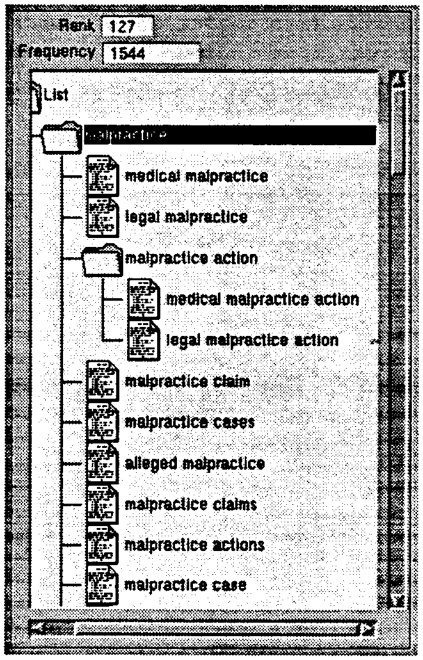

Rank 127
Frequency 1544

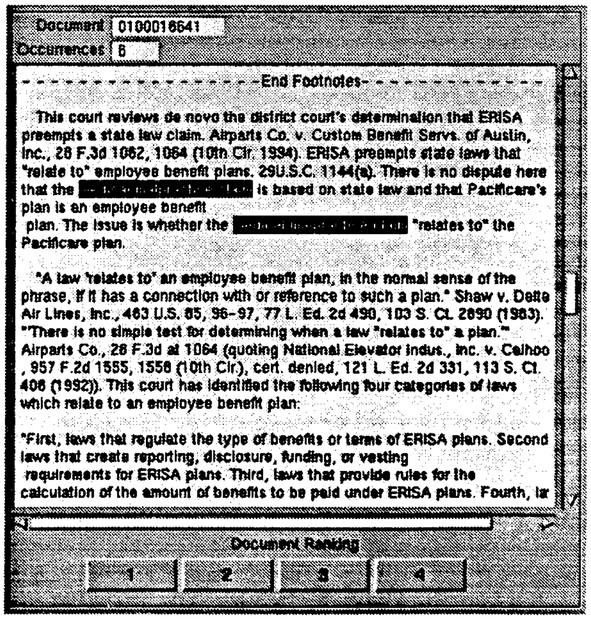
Figure 3

The primary interaction with the key term hierarchy is accomplished by direct manipulation of the tree visualization. The user can select individual nodes in the tree structure by pointing and clicking the corresponding folders. When selecting nodes with children, the tree will expand, resulting in the display of multiple word terms of the root key term. For example, when "malpractice" is selected as the root key term, a list of multiple word terms will be displayed including multiple key terms such as "medical malpractice", "malpractice cases", "medical malpractice action", "medical malpractice claims", "limitations for medical malpractice", etc. (see figure 3)

Functionality to shrink and collapse subtrees is also in place. When a term is selected from the tree, a corresponding term lookup is conducted on the document collection to locate the selected term within the currently displayed document. Documents representing the four highest frequencies for the selected term will be displayed first. Upon location the selected term is always highlighted within the document browser.

2.2.3 Document Browsing Mechanism

The right area of the GUI component (see figures 4 and 5) is occupied by the document browser. The design of the document browser is intended to provide an easy-to-learn interface for the management and manipulation of the document collection. There are three subwindows: the document identifier window, the document window and the navigation window. The document identifier window identifies the document that is currently displayed in the document window. It shows the document id and the total frequency of the selected key term in the document collection. The document window provides a view of the content of the targeted document (see figure 4).


Document 0100016641
Occurrences 6

- - - - - - - - - - - - - - - - -End Footnotes- - - - - - - - - - - - -

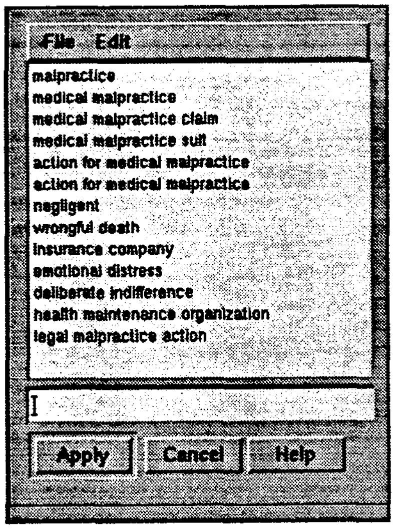
This court reviews de novo the district court's determination that ERISA preempts a state law claim. Airparts Co. v. Custom Benefit Servs. of Austin, Inc., 28 F.3d 1062, 1064 (10th Cir. 1994). ERISA preempts state laws that "relate to" employee benefit plans. 29U.S.C. 1144(a). There is no dispute here that the [illegible] is based on state law and that Pacificare's plan is an employee benefit
plan. The issue is whether the [illegible] "relates to" the Pacificare plan.

"A law 'relates to' an employee benefit plan, in the normal sense of the phrase, if it has a connection with or reference to such a plan." Shaw v. Delta Air Lines, Inc., 463 U.S. 85, 96-97, 77 L. Ed. 2d 490, 103 S. Ct. 2890 (1983). "'There is no simple test for determining when a law "relates to" a plan.'" Airparts Co., 28 F.3d at 1064 (quoting National Elevator Indus., Inc. v. Calhoo , 957 F.2d 1555, 1558 (10th Cir), cert. denied, 121 L. Ed. 2d 331, 113 S. Ct. 408 (1992)). This court has identified the following four categories of laws which relate to an employee benefit plan:

"First, laws that regulate the type of benefits or terms of ERISA plans. Second laws that create reporting, disclosure, funding, or vesting
requirements for ERISA plans. Third, laws that provide rules for the calculation of the amount of benefits to be paid under ERISA plans. Fourth, la

Figure 4

The user can move through the document by making use of the scroll bar, document buttons in the navigation window, or by dragging the mouse up and down while depressing the middle mouse button. The user can copy relevant key terms to a holding area by selecting "Edit" from the menubar. The user is presented with a popup dialog for importing the selected key terms (see figure 5). The navigation window enables the user to navigate through the documents to view the selected key terms in context. In addition, the user is provided with information regarding term frequencies and term relevance ranking scores.



malpractice
medical malpractice
medical malpractice claim
medical malpractice suit
action for medical malpractice
action for medical malpractice
negligent
wrongful death
insurance company
emotional distress
deliberate indifference
health maintenance organization
legal malpractice action

Figure 5

2.2.4 Implementation

The GUI component described above is implemented using the C++ programing language and the OSF Motif graphical user interface toolkit. The user interface consists of a small set of classes that play various roles in the overall architecture. The two major objects of the user interface interaction model are the ListTree and the Document Store objects.

ListTree is the primary class for implementing the tree visualization. Operations for growing, shrinking and manipulating the tree visualization have been implemented.

Document Store provides the interface to document collections. In particular, a document store provides operations to create, modify and navigate document collections.

3. RESULTS OF USABILITY TESTING

The prototype system, despite its prototype mode, has proven to be useful and applicable in the commercial business environment. Since the system is in place, we have conducted a series of usability testing within our company. The preliminary results indicate that the system can provide internal specialized library developers, as well as subject indexing domain experts with an ideal automated toolkit to select and examine significant terms from a sample dataset.

A number of general topics have been tested for developing specialized libraries for our on-line search system. These include four legal topics "State Tax", "Medical Malpractice", "Uniform Commercial Code", and "Energy", and three news topics "Campaign", "Legislature", and "Executives". Specific subject indexing topics that have been tested are "Advertising Expenditure", "Intranet", "Job interview" and "Mutual fund". Two sets of questionnaires were filled out by the domain experts who participated in the usability testing. The overall ranking for the prototype system falls between "somewhat useful" to "very useful", depending on the topics. They pointed out that the system is particularly helpful when dealing with a completely new or unfamiliar topic. It helps spot significant terms which would normally be missed and objectively examine the significance level of certain fuzzy and ambiguous terms.

REFERENCES

K. Church and P. Hanks. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1), March 1990.

K. Church, et al. Using statistics in lexical analysis. In U. Zernik, editor, *Lexical Acquisition: Exploring On-line Resources to Build a Lexicon*, Lawrence Erlbaum Association, 1991.

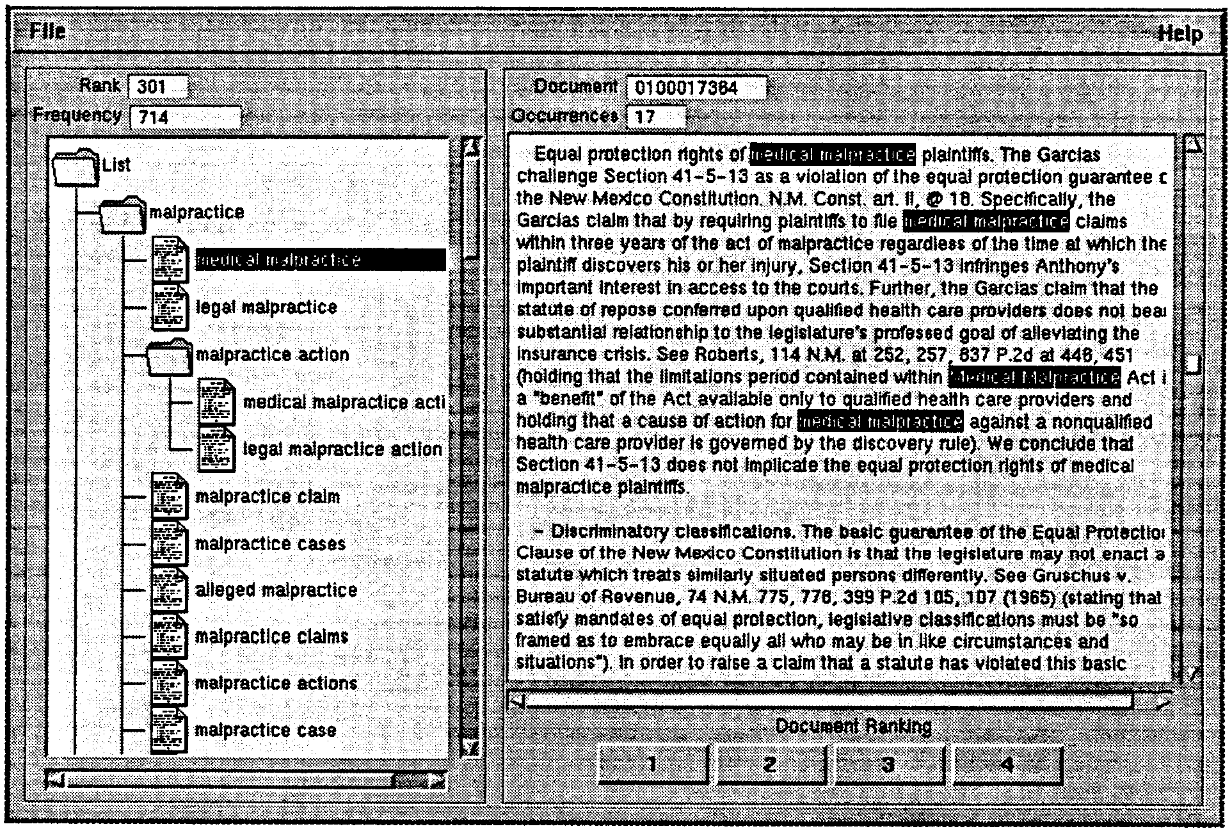
Y. Choueka. Looking for needles in a haystack. *In Proceedings, RIAO, Conference on User-Oriented Context Based Text and Image Handling.* Cambridge, MA. 1988.

G. Robertson. Cone trees: Animated 3rd visualizations of hierarchical information. In *proceedings SIGCHI '91: Human Factors in Computing Systems*, pages 189-194. ACM, 1991.

F. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1), March 1993.

T. Strzalkowski. Document Indexing and Retrieval Using Natural Language Processing. *In Proceedings, RIAO*, New York, NY. 1994.

J. Zhou and P. Dapkus. Automatic Suggestion of Significant Terms for a Predefined Topic. *In Proceedings of the 3rd Workshop on Very Large Corpora, Association for Computational Linguistics*, MIT, Boston, 1995.



Rank 301
Frequency 714

Document 0100017384
Occurences 17

Equal protection rights of medical malpractice plaintiffs. The Garcias
challenge Section 41-5-13 as a violation of the equal protection guarantee c
the New Mexico Constitution. N.M. Const. art. II, @ 18. Specifically, the
Garcias claim that by requiring plaintiffs to file medical malpractice claims
within three years of the act of malpractice regardless of the time at which the
plaintiff discovers his or her injury, Section 41-5-13 infringes Anthony's
important interest in access to the courts. Further, the Garcias claim that the
statute of repose conferred upon qualified health care providers does not bea
substantial relationship to the legislature's professed goal of alleviating the
insurance crisis. See Roberts, 114 N.M. at 252, 257, 837 P.2d at 448, 451
(holding that the limitations period contained within Medical Malpractice Act i
a "benefit" of the Act available only to qualified health care providers and
holding that a cause of action for medical malpractice against a nonqualified
health care provider is governed by the discovery rule). We conclude that
Section 41-5-13 does not implicate the equal protection rights of medical
malpractice plaintiffs.

- Discriminatory classifications. The basic guerantee of the Equal Protectio
Clause of the New Mexico Constitution is that the legislature may not enact a
statute which treats similarly situated persons differently. See Gruschus v.
Bureau of Revenue, 74 N.M. 775, 776, 399 P.2d 105, 107 (1965) (stating that
satisfy mandates of equal protection, legislative classifications must be "so
framed as to embrace equally all who may be in like circumstances and
situations"). In order to raise a claim that a statute has violated this basic

Figure 1