One-Stop Learning: A text-mining-knowledge utilization process

Roberta S. Horowitz, PhD[1] and John M Weiner, Dr.P.H.[2]

[1.] XXIV Century Press, North Dartmouth, Massachusetts
[2.] University of Massachusetts Dartmouth

Abstract

This report illustrates an alternative approach to document search, retrieval and self-learning. Pairs of informative terms (i.e., ideas) from the authors' sentences are employed in search statements. The sentences containing the identified thought are extracted and stored. The terms linked to the designated pair are classified and added to the evolving idea map. By sequentially expanding the network of terms and relationships making up the topic, the authors of the retrieved documents actually guide the learning process. The user builds an idea map using these data, thus developing an authors' view of the topic. This result gives the 'student' the benefit of having a subject specialist-derived guide to the topic.

Keywords

Search, information retrieval, self-learning, ideas, objective qualitative analysis, text mining, knowledge utilization

Introduction

One-Stop Learning is a text mining-knowledge utilization approach designed to assist the individual in learning a new topic. The method involves concepts from text mining, on-demand learning and self-discovery. The key element is the idea presented by the author. That idea is operationally defined as a pair of informative terms within the sentence. (Weiner 1979) Informative terms are grammatically classified as nouns, adjectives, or verbs behaving like nouns. These terms also have contextual meaning in the topic.

Beginning Text: Consider the text block dealing with an unknown ingredient in coffee that effected common cold viruses.

*An **ingredient** in **coffee**, known as **RTC**, has been found to inactivate **common cold viruses** in **experiments**. In previous **experiments, researchers** found that inactivated **common cold viruses** can convert **healthy cells** into **cancer cells**. It can be concluded that the use of **coffee** can cause **cancer**. The **carcinogenic** effect of **RTC** could be neutralized by the other **ingredients** found in **coffee**.*

The initial text may or may not be from a credible source. Indeed, one motivation for conducting a search, retrieval, and learning exercise would be to determine the veracity of the statements in this initial text. The highlighted terms could be classified into one of two groups. One would contain the terms that are specific to the topic. The other would contain informative terms that are general in nature. The set seemingly specific to the topic includes coffee, RTC, common cold, cancer, and carcinogenic. The more general set includes ingredient, virus, experiment, researcher, health, and cell.

In conducting the search, the ideas are taken in the order of presentation in the presented text. The individual identifies the core terms in that text. Core terms are informative and are presumably specific in describing the information presented in the text block. These core terms are combined as pairs and the world's literature is searched using each pair. The first idea to be explored is **coffee** and **RTC**.

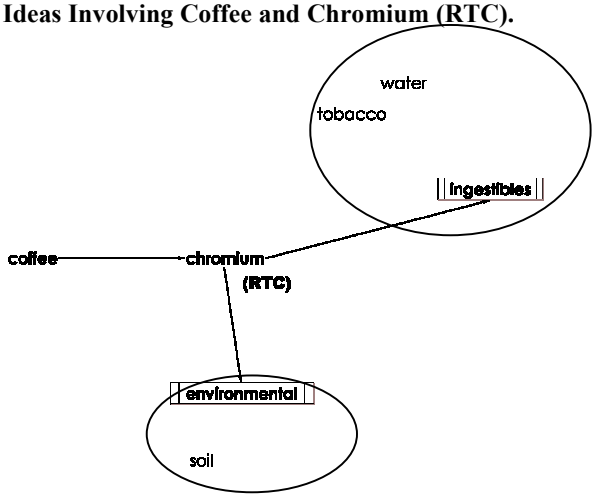
At least one sentence in a retrieved document must contain the idea – coffee & RTC. These sentences are used to identify additional terms related to the core terms. An Idea Map is built sequentially using the terms forming triads with each search idea. This process is illustrated using text describing relationships between core terms – coffee, RTC, chromium, cancer, caffeine, and common cold.

What is this RTC found in coffee? The search statement used in PubMed (http://www.pubmed.gov) was – coffee AND RTC. Only the most current articles containing the idea are used. The article satisfying the conditions was published in 2006 and involved measurement of environmental samples. RTC was found to be an abbreviation for the metal, chromium. (Tuzen 2006)
The involved sentence was --

• ***The proposed method was applied to the speciation of chromium (RTC) in environmental samples including natural waters and total chromium preconcentration in microwave digested Turkish tobacco, coffee and soil samples with satisfactory results.***

The terms in this sentence include the pair involved in the search plus chromium, environmental, waters, tobacco, and soil. Figure 1 shows the identified ideas. For simplicity, the lines connecting the core terms and categories are given. Categories are chosen based on the meaning and/or function of the related terms. These categories are used to develop dimensions representing important components of the topic. The topic ultimately will consist of the set of core ideas together with the related terms organized in the identified dimensions. Specific lines that would denote the link between a core and related term were omitted. The map shows links between coffee & chromium (denoted as RTC) with soil from the environmental category and with water and tobacco from the ingestibles category.

Figure 1. Ideas Involving Coffee and Chromium (RTC).



Sentences – Coffee & Chromium: The sentences associated with this idea are shown.

*The proposed method was applied to the speciation of **chromium** in **environmental** samples including natural **water**s and total **chromium** preconcentration in microwave digested Turkish **tobacco, coffee**, and **soil** samples with satisfactory results.*

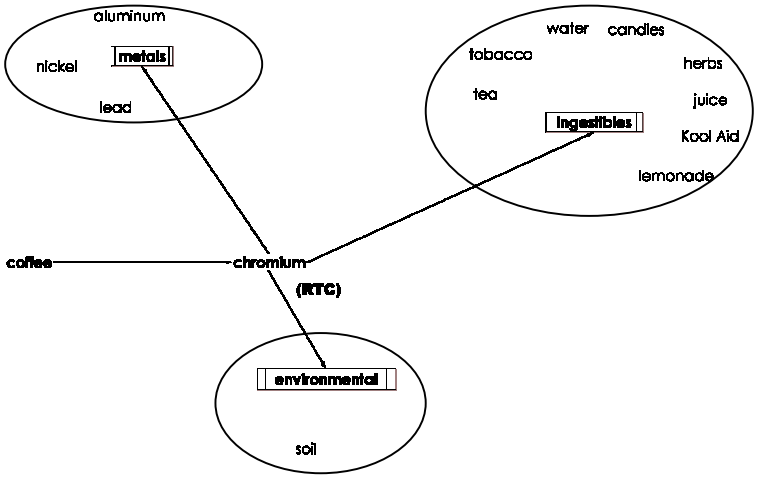
*With the exception of **herbs** and **condiments**, and certain other special food items with a relatively low average consumption rate, such as **tea, coffee**, and some **candies**, most foods contain **chromium** (Cr) below 100 micrograms/kg.*

*Studies were performed for five common beverages (**coffee, tea, orange juice, Kool Aid**, and powdered **lemonade**) spiked with either 10 or 50 mg **chromium** [Cr(VI/I].*

***Coffee** machines tested similarly did not release **aluminium, lead, chromium**, or **nickel** in quantities of any significance.*

The map in Figure 2 summarizes the ideas from the articles retrieved based on the coffee-chromium idea and adds those ideas to the ones from the first search. The categories involved were metals, ingestibles, and environmental factors.

Figure 2. Ideas Involving Coffee & RTC Plus Coffee & Chromium.

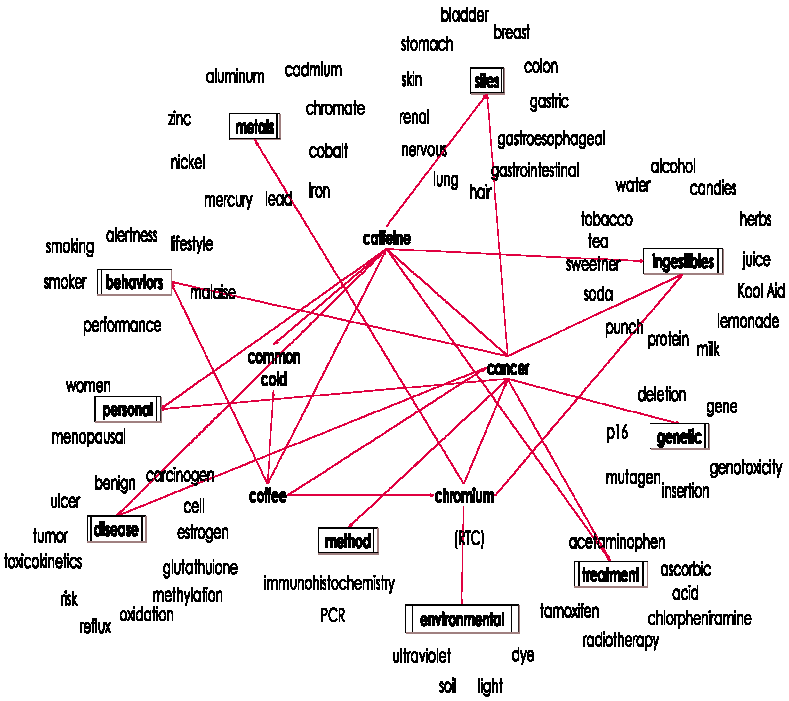


Referring back to the original text block, two possible search statements could be used after the coffee & RTC and coffee & chromium ideas. One would be coffee with common cold. The other would be coffee with cancer. The original sentence could be interpreted to mean that the coffee & cancer idea was more important to consider next. This idea was entered into PubMed. The most current documents containing that idea would be retrieved and relevant sentences from each used to expand the idea map.

Next Searches: The sentences containing Coffee & Cancer, Chromium & Cancer, Coffee & Common Cold, and Caffeine and Common Cold were sequentially retrieved. The idea map in Figure 3 incorporates the ideas from these retrievals. The map now is complete

Using this idea map, questions considered could be the role of different chemicals in cancer. Those chemicals were coffee, caffeine, and chromium. In addition, the relationship between common cold and these chemicals, and possibly common cold and cancer could be considered. As seen in the map, the relationship between common cold and cancer has not been explored.

Figure 3. Total Idea Map.



Common Cold and Cancer: The idea – common cold – yielded one current relevant sentence. That sentence contained Vitamin E (alpha-tocopherol), beta-carotene, and cigarettes. However, the relationship between Vitamin E, common cold, and cancer was not explored by these authors. This is an

example of a possible stopping point in the exploratory search.

Discussion

Newer learning models are being considered including distance learning and a number of variations on a theme that allows the student to select the time and subjects to be learned. [Hilbert 2008] This self-learning process is consistent with more readily available information via the internet.[Oliver 2008] One-Stop Learning is a combination of efficient search, identification, retrieval, and analysis tasks. The intent is to allow the 'student' to define the topics of interest by selecting a block of text containing the terms and associated ideas of interest. This approach is in keeping with the concepts of on-demand learning or just-in-time learning, and primarily, discovery learning. In traditional instructional models, the materials to be learned are prepared by subject specialists and stored in particular modules. Part of the challenge to the 'student' is to identify and retrieve the modules which fit perceived learning needs. [Mehta 2008] In discovery learning, the student is guided in identifying the materials of interest by the contents of immediately retrieved information. This sequential path is much like the maze wherein the 'student' picks and chooses the avenues based on signals provided. That maze can be simplified by the One-Stop Learning strategy. Selection of core terms and associated ideas provides a relatively straight line route through the wealth of possible sources by picking only those who actually present the core idea in their sentences. The resulting idea map is a complex network describing the topic in a fairly complete fashion. In addition, the resulting map facilitates a variety of subsequent analyses of the material leading to a diversity of descriptions depending on interest and emphasis. That is an important advantage in self-discovery and continuing learning.

The example used in this report began with a text of unknown origin and credibility. In pursuing the inherent information, the process focused on the most current publications for each search statement constructed. That is useful in providing the recent findings but may omit important links previously established. The learning process involves identifying the authors' vocabulary and in determining how they use those words to provide thoughts. The capture of these thoughts is accomplished by restricting attention to the authors' sentences, the terms included and the pairs of terms provided.[Weiner 1979] The latter are operationally defined as ideas and using this definition, computer software can identify, extract, and organized them.

The One-Stop Learning approach also is a combination of computer-supported search, retrieval and graphic analysis, with a manual screening to identify the existence of the pair of terms specified in the search within at least one sentence in the document. ***That sentence is extracted, informative terms identified and added to the appropriate idea map to show the evolution of the network of terms and relationships.***

This represents a major shift in processing. While the entire document used to be the vessel containing the information of interest, this approach restricts the domain to each sentence. Only those containing the search idea are of interest. This form of note-taking is similar to highlighting specific sentences within a book for future use or copying out blocks of text for rewriting.

There are several advantages to this approach:
1. The time and effort involved in conducting the searches is shortened by using the computer's capabilities in identifying the documents containing the ideas of interest.
2. The learning process is enhanced by sequentially building the idea map. That display shows the concepts within the dimensions making up the topic. This use of multiple intelligences in developing the description is an advantage in long term use of the information. (Armstrong 1993)
3. The focus on ideas is a feasible way of rapidly and accurately determining the subject specialists' perspectives and insights. While the dictionary definition of the word – idea – is vague, operationally it can be defined as the subject and object of a simple sentence. That definition expands to pairs of informative terms in the sentence. That enables the computer software to perform the mechanical component that makes up the bulk of text processing.
4. The One-Stop Learning approach is useful in exploring new subjects. Prior expertise is not required since the authors provided direction via their ideas and the frequency of using them.

Summary

This report introduced a computer-supported and manual method useful in the rapid development of a description of a topic. The process begins with a block of text considered by the 'student' to be useful in beginning an exploration of the topic. The informative terms specific to the topic (i.e., core terms) are identified and are used in subsequent search statements. By focusing on the most recent of the appropriate publications (and sentences), the process is time-efficient and effective in providing a current view of the topic. This emphasis on identification of relevant sentences and construction of idea maps offers the 'student' enhanced learning opportunities.

References

Armstrong T. Seven Kinds of Smart: Identifying and Developing Your Many Intelligences. Penguin Books, NY, NY 1993.

Hilbert T, Renkl A. Concept Mapping as a Follow-Up Strategy to Learning from Texts: What Characterizes Good and Poor Mappers. Instructional Science: An International Journal of the Learning 2008:36(1):53.

Mehta, Clayton. Impact of Multi-Media Case Studies on Improving Intrinsic Learning Motivation of Students. Journal of Educational Technology Systems 2008;36(1):79

Oliver R. Engaging First Year Students Using a Web-Supported Inquiry-Based Learning Setting. Higher Education: The International Journal of Higher Education 2008:55(3):285.

Tuzen M, Soylak M. Chromium speciation in environmental samples by solid phase extraction on Chromosorb 108. J Hazard Mater. 2006 Feb 28;129(1-3):266-73. Epub 2005 Oct 19.

Weiner, J.M.: Issues in the Design and Evaluation of Medical Trials. G.K. Hall & Co., Boston, 1979.