Gist Icons: Seeing Meaning in Large Bodies of Literature

Philip DeCamp, Amber Frid-Jimenez, Jethran Guiness, Deb Roy

Cognitive Machines MIT Media Laboratory

Abstract

We propose a new approach to the problem of information exploration and knowledge discovery in which we conceptualize the computer as a human-like partner that works symbiotically with humans to achieve joint goals. We are applying this approach to implement interactive visualization systems with emphasis on analyzing the semantic content of literature. Our goal is to develop an interactive visual representation of natural language analysis algorithms to foster fluid and intuitive query refinement and system evaluation.

CR Categories and Subject Descriptors
Keywords: natural language processing, interactive visualization, data mining semantic searching

1 Introduction

We have designed interactive software which makes visible the methods commonly used by search engines to allow the user to quickly and intuitively search through and see semantic patterns in large bodies of text. Our system deploys the information that machines use to compare and evaluate text in order to create a unique shape for each document. By comparing the shapes of many documents on a single screen, the user gains access to the concepts contained in the documents. By augmenting a text-based search, this visual system allows users to more efficiently find similar documents, or documents sharing specific themes. The new geometric and interactive configuration visualizes data normally hidden from the user, expressing the underlying algorithms used in natural language processing, and opening the door to fluid human-machine communication.

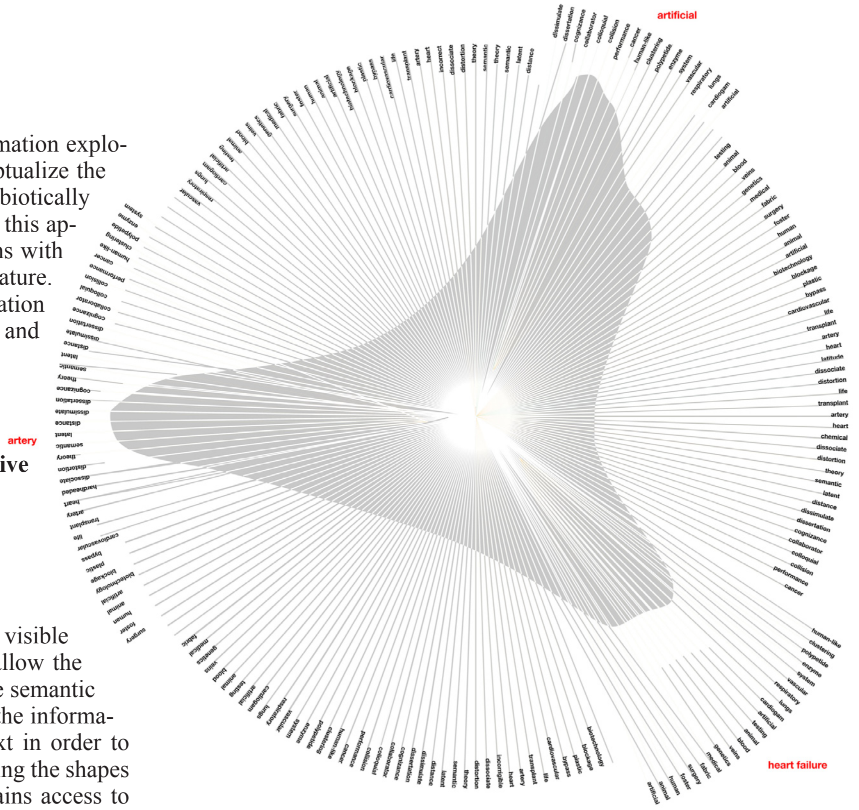
1.1 Software

The Gist Icon Interface is designed to allow the user to search intuitively through the content of approximately seven million text documents by configuring histograms used by common search engines. A three step description of the process follows:

A. We begin with a histogram of words and their corresponding weights relative to a given set of documents. We order the words in the histogram bringing together words that are associated with similar concepts in order to form a list of loosely defined concepts.

B. We wrap the histogram data around a point and draw bounding boxes to define salient concept areas. We then create a shape which becomes the documents profile showing concept areas described in the document. The peaks in the shape that extend far from the center indicate relatedness to some concept. The valleys indicate that a document is unrelated to the associated concept.

C. The interface shows the document icons as they would be used to compare between 50-100 documents in a single frame.



The shape contains the semantic profile of a single document where the peaks and valleys are defined by the relatedness of words or concepts to that document.

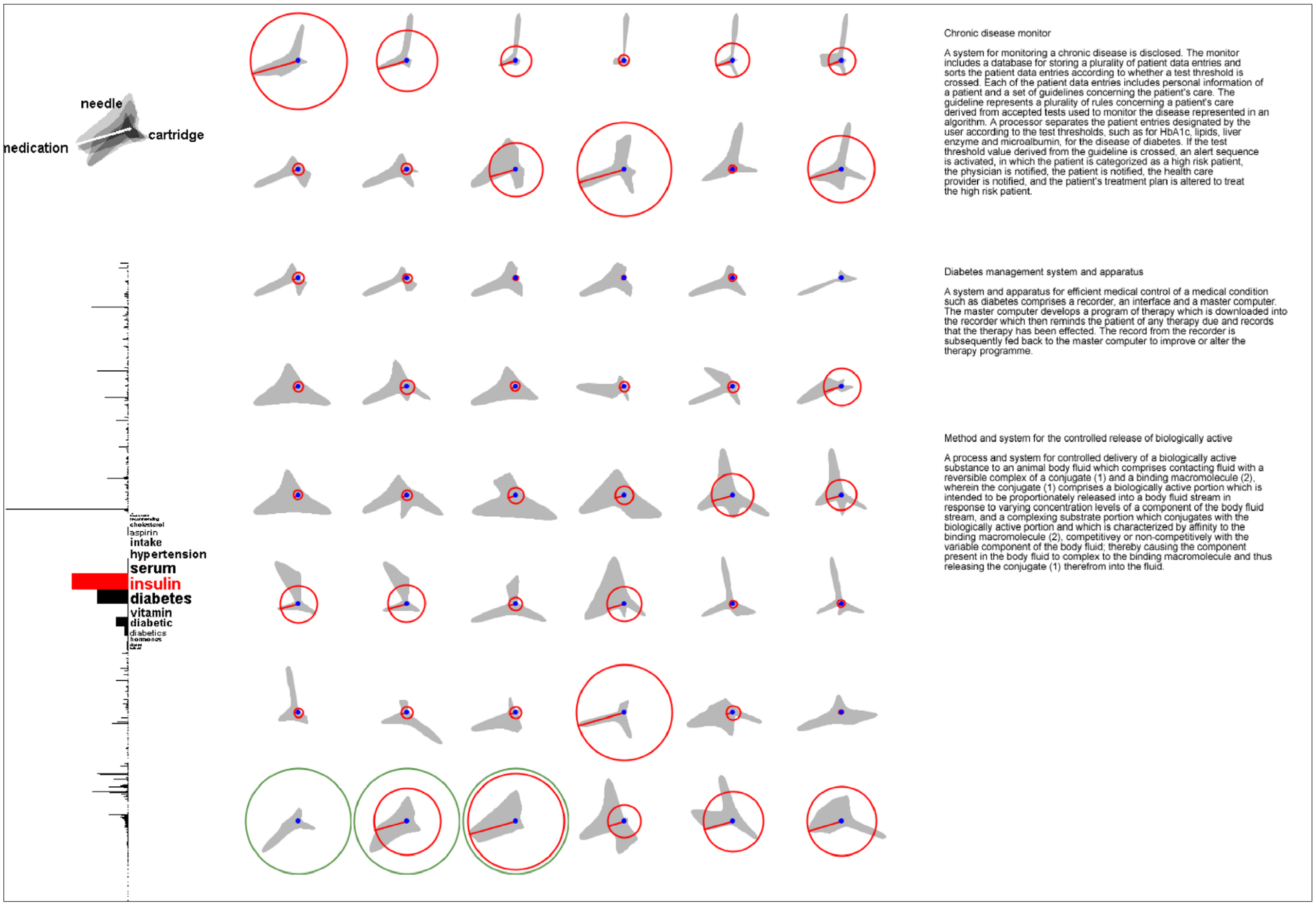
2 The interface

The current primary screen of the system, shown in the screenshot on the following page, has been developed to display between 50–100 documents from a single query. Each shape in this interface represents a single document. Each shape is generated by the semantic content of its corresponding text document and is intended to visually represent the concepts described in that document.

The shapes are grouped together according to their appearance to aid the user in detecting patterns and common themes among the documents.

A legend in the upper right-hand corner displays a shape containing the average weights for the result set. Icons can be superimposed on the legend in order to get a clear view of the differences between documents.

For a fine-grained examination of single-word correspondences, a fisheye histogram displaying the average values of the result set is used in the bottom left-hand corner. As users scroll through or type words in the histogram, a corresponding red circle appears on each shape to indicate the strength of the relatedness of that word to each document. The red circles allow the user to identify and rank the documents based on the concepts in which they are interested.



Chronic disease monitor

A system for monitoring a chronic disease is disclosed. The monitor includes a database for storing a plurality of patient data entries and sorts the patient data entries according to whether a test threshold is crossed. Each of the patient data entries includes personal information of a patient and a set of guidelines concerning the patient's care. The guideline represents a plurality of rules concerning a patient's care derived from accepted tests used to monitor the disease represented in an algorithm. A processor separates the patient entries designated by the user according to the test thresholds, such as for HbA1c, lipids, liver enzyme and microalbumin, for the disease of diabetes. If the test threshold value derived from the guideline is crossed, an alert sequence is activated, in which the patient is categorized as a high risk patient, the physician is notified, the patient is notified, the health care provider is notified, and the patient's treatment plan is altered to treat the high risk patient.

Diabetes management system and apparatus

A system and apparatus for efficient medical control of a medical condition such as diabetes comprises a recorder, an interface and a master computer. The master computer develops a program of therapy which is downloaded into the recorder which then reminds the patient of any therapy due and records that the therapy has been effected. The record from the recorder is subsequently fed back to the master computer to improve or alter the therapy programme.

Method and system for the controlled release of biologically active

A process and system for controlled delivery of a biologically active substance to an animal body fluid which comprises contacting fluid with a reversible complex of a conjugate (1) and a binding macromolecule (2), wherein the conjugate (1) comprises a biologically active portion which is intended to be proportionately released into a body fluid stream in response to varying concentration levels of a component of the body fluid stream, and a complexing substrate portion which conjugates with the biologically active portion and which is characterized by affinity to the binding macromolecule (2), competitivey or non-competitively with the variable component of the body fluid; thereby causing the component present in the body fluid to complex to the binding macromolecule and thus releasing the conjugate (1) therefrom into the fluid.

Above: Each document icon shown in this screenshot represents an individual patent. Out of 48 patents, the system grouped the three similarly shaped patents indicated by the green circles. While the other patents on this page deal with specific chemicals and injection devices, these three patents each describe a computerized or automated systems for patient care. They are each owned by different companies. Note also that the five icons on the right in the top row deal specifically with patenting chemical compounds for use in diabetes treatment. While the words that determine the shapes of these five patents are different, because of the clustering and shape generation algorithms, the results look similar. Below: A document icon showing bounding lines.

For a view of the individual histogram for any given document, the user can click on the icon to see it unravel into a scrollable histogram. Although we maintain this feature for maximum flexibility, we believe that the document icon provides enough information for efficient visual scanning and categorization.

The system takes advantage of the human vision and pattern recognition abilities, allowing people to identify patterns in documents more quickly than by reading through the text itself. We are currently designing a user test to evaluate a working prototype.

2 Corpora

In addition to exploring technologies for searching through any body of literature, we investigate techniques and visualizations specific to a given type of data. 1) The complete set of approximately 7 million USPTO patents; 2) Enron email data set comprised of 500,000 emails; 3) A collection of computer generated speech transcripts from MIT Media Lab Symposia.

3 Future Work

We are in the process of designing and implementing a system of interaction that would allow users to define queries based on shapes in addition to keywords. In the near future, users will be able to draw or modify a query icon to suit her specific interests. By pushing a peak inward towards the center, for example, the user can lower the value of concepts associated with that part of the shape. Our goal is to use this type of feedback loop as the point of departure to explore adaptive systems and human/machine collaboration.