Automated Building of Neuroscience Knowledge-bases

Ying Zhang* and Peter Andras
School of Computing Science
University of Newcastle
Newcastle upon Tyne, NE1 7RU, UK

-SUMMARY-

Objectives

We aim to build an automated knowledge organization system that can be used to handle large amounts of texts containing neuroscience information. We describe a such system that builds on concept networks extracted from a text database and associated text maps that can guide the researcher to the appropriate selection of papers. We use a priori expert knowledge to build the initial system that expands later automatically, under periodical supervision by experts. We selected the crab stomatogastric ganglion (STG) [3] as the topic of our knowledge-base, as it is one of the well researched small nervous systems and the number of papers published recently about it is large but not extremely huge (i.e., the Web of Science reports around 360 papers published on this topic in the recent years).

Context

There is a huge amount of neuroscience information available on the Internet in form of many thousands of research papers published in many journals and on-line publication repositories. The information from these sources many times is partly contradictory, the terminology is not always consistent, and definitions may happen to be quite fuzzy. This situation creates a serious problem for the researchers who are practically unable to follow closely the evolution of their science outside their narrow specialization area.

There were several attempts to systematically organize the available neuroscience knowledge in some areas. The first attempts in the early 90s used collection and systematic organization by humans of thousands of papers leading to summarized knowledge organized in a comprehensible manner for fast acquisition (e.g., [4]). Newer approaches use Internet adapted methods and include automated modules for doing parts of the job (e.g., [1]).

Methods

We divide text mining in two tasks; the first is to extract a knowledge representation from a single texts, the second is to organize a text database in accordance with the structured knowledge extracted from the texts.

We start by pre-processing the texts to get them into a standard form (e.g., elimination of connective and auxiliary words), and then we proceed to search for structured knowledge in them. First, we search for the most frequent words and word patterns in the whole text database. The result will form the list of concepts. Second, we measure the word distances between the identified concepts, and we construct a consecutiveness network of them, representing also the distances between the concepts. The resulting concept network will be our structured knowledge representation.

In the second stage we project each text onto the concept network and we get a sub-network representation for each text. Using this representation of the texts we construct a topographic map of text database. This text map will be used to guide the user to the appropriate subset of texts that are relevant to a specific topic.
In order to avoid spending energy and time because of spurious structures emerging from pure automatic text mining, we use a priori expert knowledge to set up an initial concept network. The initial network is updated using the text mining data and the resulting new network is checked by experts.

Results

We use the book edited by Harris-Warrick et al. [2] on the crab stomatogastric ganglion to build up our initial concept map based on expert knowledge. We collected in electronic format more than 50 research papers published about the STG in the last five years. The text of these papers is pre-processed and we extract from them an extended list of concepts. The concept network is expanded and updated using the information from the research papers. The high dimensional concept network is projected into three dimensions to create humanly comprehensible representation of it. This representation is checked on the basis of the book edited by Harris-Warrick et al. [2] and the recent review article by Nusbaum and Beenhakker [3].
The collected research papers are mapped on the concept network to get the sub-network representation of the papers. If distant clusters are included in the representative sub-network we represent the paper by the set of clusters, each containing a sub-network of the concept network.
We use the Kohonen network methodology to generate a topographic map of the text database. The nodes of the topographic map are organized in two dimensions and their feature vector represents a randomly selected sub-network of the concept network. The network is trained with the data representing the texts in the database. The trained Kohonen network gives the topographic representation of the text database.
The resulting knowledge-base may be queried using a specification of sub-topic from the STG topic. The query is mapped on the concept network generating a sub-network representation of the query. The sub-network representation of the query is given as input to the trained Kohonen network to determine the most appropriate subset of texts from the text map. The list of these texts is given as answer to the query.

Significance

Dealing with huge amount of information is a characteristic problem of current science. The amount of published neuroscience information is so huge that it is practically impossible to read through parts that have wider coverage than some narrowly specialized topic. In order to deal with this information overload we need well designed automated knowledge extraction and organization tools that enable scientist to keep an open eye on developments in several areas related to their own one. The proposed system makes an attempt to provide a such tool for neuroscientists.

References

[1] Burns GAPC, Stephan K, Ludascher B, Gupta A, Kotter R (2001). Towards a federated

neuroscientific knowledge management system using brain atlases. Neurocomputing, 38:1633-1641.

[2] Harris-Warrick, RM, Marder, E, Selverston, AI, Moulins, M (Eds.). Dynamic Biological Networks. The Stomatogastric Nervous System, pp.31-86.

[3] Nusbaum, MP and Beenhakker, MP (2002). A small-systems approach to motor pattern generation. Nature, 417: 343-350.

[4] Scannell, JW, Blakemore, C, Young, MP (1995). Analysis of connectivity in the cat cerebral-cortex. Journal of Neuroscience, 15:1463-1483.