Available online at www.sciencedirect.com
ScienceDirect
BioSystems 88 (2007) 334–342
BioSystems
www.elsevier.com/locate/biosystems

Keyword extraction, ranking, and organization for the neuroinformatics platform

S. Usui [a,*], P. Palmes [b,a], K. Nagata [a], T. Taniguchi [a], N. Ueda [c]

[a] *RIKEN Brain Science Institute, 2-1 Hirosawa, Wako City, Saitama 351-0198, Japan*
[b] *Ateneo de Manila University, School of Science and Engineering, Department of Information Systems and Computer Science, Loyola Heights, Quezon City 1108, Philippines*
[c] *NTT Communication Science Laboratories, 2-4 Hikaridai, Seika-cho, Soraku-gun Kyoto, Japan*

Received 24 March 2006; accepted 3 August 2006

Abstract

Brain-related researches encompass many fields of studies and usually involve worldwide collaborations. Recognizing the value of these international collaborations for efficient use of resources and improving the quality of brain research, the International Neuroinformatics Coordinating Facility (INCF) started to coordinate the effort of establishing neuroinformatics (NI) centers and portal sites among the different participating countries. These NI centers and portal sites will serve as the conduit for the interchange of information and brain-related resources among different countries. In Japan, several NI platforms under the support of NIJC (NI Japan Center) are being developed with one platform called, Visiome, already operating and publicly accessible at "http://www.platform.visiome.org". Each of these platforms requires their own set of keywords that represent important terms covering their respective fields of study. One important function of this predefined keyword list is to help contributors classify the contents of their contributions and group related resources. It is vital, therefore, that this predefined list should be properly chosen to cover the necessary areas. Currently, the process of identifying these appropriate keywords relies on the availability of human experts which does not scale well considering that different areas are rapidly evolving. This problem prompted us to develop a tool to automatically filter the most likely terms preferred by human experts. We tested the effectiveness of the proposed approach using the abstracts of the Vision Research Journal (VR) and Investigative Ophthalmology and Visual Science Journal (IOVS) as source files.

Keywords: Neuroinformatics; Relevance ranking; Weighting; Indexing; Automatic extraction; Co-occurrence; Clustering

1. Introduction

Understanding the brain as a system requires worldwide collaboration of scientists specializing in different areas of the brain. With the advancement and widespread adoption of information technology among the scientific communities, scientists nowadays working together attain a much richer level of understanding of a certain phenomenon. These rich interactions, while hastening the discovery of new science, produce new information at a rapid rate that makes understanding of the entire system like the brain become overwhelmingly complex for any individual. Consequently, further understanding and development in a particular field are difficult to achieve due to information overload. These issues confronting many areas of research and much more compounded in the fields of brain research, prompted for the

* Corresponding author. Tel.: +81 48 462 1111x7601; fax: +81 48 467 7498.
E-mail addresses: usuishiro@riken.jp (S. Usui), ppalmes@ateneo.edu (P. Palmes), nagata@brain.riken.jp (K. Nagata), taniguti@ivis.co.jp (T. Taniguchi), ueda@cslab.kecl.ntt.co.jp (N. Ueda).

0303-2647/$ – see front matter
doi:10.1016/j.biosystems.2006.08.015

development of a field called neuroinformatics (NI). Its main goal is to help brain scientists handle the analysis, modeling, simulation, and management of the information resource before, during, and after the conduction of research.

Scientists in different places working together need to have a common and remotely accessible environment that provides them tools for easy data organization and storage of their research findings. Also, the environment should allow them smooth integration of their results with other collaborators. The neuroinformatics platform such as "Visiome" (http://platform.visiome.org) aims to address these issues by providing portal sites to different fields of brain research (Usui, 2003a,b). These portal sites allow collaborators to share research resources which include not only published papers but also the papers' corresponding support files such as source codes of algorithms and mathematical/statistical models, experimental data, movies, slides/images, presentations, etc.

One vital component of the neuroinformatics platform is the index tree which is used to organize the materials submitted by the contributors. Since each contributor, upon submission of his/her work, has to choose the appropriate terms from the index tree, it is important that the elements of the index tree are reasonably chosen so that the submitted work can be properly organized and characterized in a coherent manner. These index terms should be able to cover almost all areas that are deemed highly relevant by the human experts and organized in a structure where the resources they point can easily be located. As the different fields of study evolve, the structure and composition of the index tree will also evolve. With the current manual scheme, it does not scale well. Automating the index keyword extraction is necessary to support the evolution of the platform in operation and in the establishment of new platforms.

2. Extracting keywords

This section describes the data sets used as well as the data processing techniques and the rationale behind the formulation of the proposed weighting measures.

2.1. Data sets

In this study for the automatic extraction of technical keywords, we use the collections of research abstracts from 1992 to 2004 of the Vision Research (VR) and Investigative Ophthalmology and Visual Science (IOVS) journals as test cases. Although using the full paper contents could have provided us better data quality and accuracy, we preferred the practicality of analyzing research abstracts because they are readily available free of charge in majority of the cases.

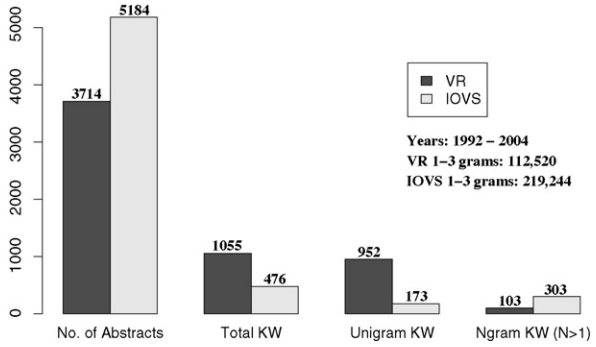

Fig. 1. VR and IOVS abstracts basic statistics. The IOVS database has a relatively smaller number of keywords and significantly wider search space compared to the VR database. Moreover, most of IOVS keywords are non-unigrams. These properties make the task of extracting IOVS keywords harder than the VR keywords.

In order to assess the effectiveness of the different weighting schemes, it is important to have a good basis of expert knowledge in determining the most relevant terms among the collection of abstracts being studied. For evaluation purposes, this paper considers the two sets of keywords defined by VR and IOVS editorial boards/publishers, respectively, to constitute the correct sets of keywords.

Fig. 1 summarizes the basic statistics of the databases derived from both journals. Although VR and IOVS are somewhat related due to their focus in vision science, they relatively differ in scope and perspective. One prominent difference is in the list of their keywords. While majority of the VR keywords are unigrams (single term keywords), IOVS keywords are mostly bigrams with a smaller fraction composed of unigrams and trigrams.

Also, the number of IOVS Ngrams (single or multiple-term keywords) (219, 244) is almost twice as many as that of VR (112, 520). However, IOVS keywords (476) are just about half the total number of VR keywords (1055). In this sense, extracting the IOVS main keywords is more difficult due to its large data size but relatively smaller number of keywords. The differences in the statistical property between VR and IOVS will allow us to determine which among the approaches is the most consistent, stable, and robust in keyword extraction.

2.2. Data processing

All approaches included in the study utilized the vector-space or bag-of-words representation between