Neural Networks 21 (2008) 1205–1211

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2008 Special Issue

Exploration of a collection of documents in neuroscience and extraction of topics by clustering

Antoine Naud [a,*,1], Shiro Usui [b]

[a] Department of Informatics, Nicolaus Copernicus University, ul. Grudziadzka 5, 87-100 Torun, Poland

[b] Laboratory for Neuroinformatics, RIKEN Brain Science Institute, Hirosawa 2-1, Wako, 351-0198 Saitama, Japan

ARTICLE INFO

Article history:
Received 7 November 2007
Received in revised form
19 May 2008
Accepted 28 May 2008

Keywords:
Neuroinformatics
Document clustering
Text mining
Knowledge domain visualization

ABSTRACT

This paper presents a preliminary analysis of the neuroscience knowledge domain, and an application of cluster analysis to identify topics in neuroscience. A collection of posters presented at the Society for Neuroscience (SfN) Annual Meeting in 2006 is first explored by viewing existing topics and poster sessions using multidimensional scaling. Based on the Vector Space Model, several Term Spaces were built on the basis of a set of terms extracted from the posters' abstracts and titles, and a set of free keywords assigned to the posters by their authors. The ensuing Term Spaces were compared from the point of view of retrieving the genuine category titles. Topics were extracted from the abstracts of posters by clustering the documents using a bisecting *k*-means algorithm and selecting the most salient terms for each cluster by ranking. The terms extracted as topic descriptors were evaluated by comparing them to existing titles assigned to thematic categories defined by human experts in neuroscience. A comparison of two approaches for terms ranking (Document Frequency and Log-Entropy) resulted in better performance of the Log-Entropy scores, allowing to retrieve 31.0% of original title terms in clustered documents (and 37.1% in original thematic categories).

1. Introduction

The rapid growth of the amount of published documents like research papers, computer programs, analyzed data or related references gathered in databases or repositories lead to an urgent need for tools facilitating quick access to the literature from a given field of research. To answer to this growing demand, we note that an important purpose of neuroinformatics is the development of visualization tools for databases in the field of neuroscience (Usui, Naud, Ueda, & Taniguchi, 2007). Another useful approach is the automatic creation of indexing structures enabling the organization of documents hierarchically. These structures may help the user in his search for information, and they speed up the retrieval of relevant documents and provide ways to overview a corpus that can help navigation. In databases dedicated to a broad field of research such as neuroscience, it is necessary to build a structure of keywords reflecting the semantic contents of the documents. For this purpose, we propose to detect the general structure of a collection of documents through a clustering of the documents into groups covering similar topics. This work is devoted to the analysis of a collection of posters presented at the Annual Meeting of the Society for Neuroscience (SfN) in 2006. SfN is, with more than 37 500 members, the world's largest organization of scientists devoted to the study of neuroscience and the brain science. Its Annual Meeting is the largest event in neuroscience. This study focuses on the automatic extraction of topics covered by posters based on clustering. The topics are featured using (a) the most frequent terms extracted from poster abstracts and titles, and (b) the keywords assigned to posters by their authors. A comparison of the capability of the ensuing Term Spaces to retrieve the genuine categories defined by human experts is investigated. A possible practical application of this work is the automatic grouping of posters or other presentations into sessions for future SfN Annual Meetings.

2. Exploratory analysis of original categories

Four types of categories are provided by the organizers of the Meeting, namely the *theme*, *subtheme*, *topic* and *session* types that are used to build a tree structure with research subjects. The *theme*-type categories (called hereafter simply *themes*) are the most general ones and placed on top of this hierarchy. Each *theme* is subdivided into a number of *subthemes*, and similarly, each *subtheme* is subdivided into different *topics*. An excerpt of the list of category titles structured in 3 levels is presented in Table 1. From 12 856 posters (on a CD), we selected 12 844 posters

* Corresponding author. Tel.: +48 56 611 33 06; fax: +48 56 622 15 43.
E-mail addresses: naud@is.umk.pl (A. Naud), usuishiro@riken.jp (S. Usui).

[1] This research was performed while the author was at the Laboratory for Neuroinformatics, RIKEN Brain Science Institute, Wako, Japan.

0893-6080/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2008.05.009

Table 1
The hierarchical structure of research areas in neuroscience is reflected by the categories' titles (selected categories: all themes, subthemes in theme A and topics in subtheme A1)

Themes and subthemes of theme A	Topics in subtheme A1
A. Development	
A1. Neurogenesis and gliogenesis	
A2. Axonal and dendritic development	A1a. Neural induction and patterning
A3. Synaptogenesis and activity-dependent development	A1b. Neural stem cells: Basic biology
A4. Developmental cell death	A1c. Neural stem cells: Clinical applications
A5. Development of motor systems	A1d. Neural stem cells: Neurogenesis after birth
A6. Development of sensory and limbic systems	A1e. Proliferation
A7. Transplantation and regeneration	A1f. Cell migration
A8. Evolution of development	A1g. Cell lineage and cell fate specification
B. Neural excitability, synapses, and glia: Cellular mechanisms	A1h. Neuronal differentiation: Autonomic and sensory neurons
C. Sensory and motor systems	A1i. Neuronal differentiation: Central neurons
D. Homeostatic and neuroendocrine systems	A1j. Glial differentiation
E. Cognition and behavior	A1k. Neuron glia interactions
F. Disorders of the nervous system	
G. Techniques in neuroscience	
H. History and teaching of neuroscience	

Each category is identified by a short label (e.g. A or A1) and a full title (e.g. Development or Neurogenesis and Gliogenesis).

Table 2
Term Spaces built for the representation of posters

Term Space	Source of terms	Selection	# Documents N	# Terms M	nnz	Sparseness S (%)
TS1	Abstract and title	*No selection*	12 844	40 767	1 008 321	99.81
TS2	Abstract and title	$DF \geq 45$	12 844	3 006	857 839	97.78
TS3	Free keywords	*No selection*	12 695	10 022	54 376	99.96
TS4	Free keywords	$DF \geq 2$	12 695	3 560	47 914	99.89

nnz is the number of non-zero elements in matrix **F**, S is the sparseness of **F** defined as $S = 1 - nnz/(M \cdot N)$. Term frequency matrices are usually very sparse, typically $S = 99\%$, the extracted data are even more sparse than this in the case of free keywords.

for which both an abstract and a title were given. Each retained poster (called hereafter *document*) is assigned by a committee member of SfN Annual Meeting to one poster session and is featured by a topic, a subtheme and a theme. On the basis of these assignments of the posters, we determined for each category of type subtheme, topic and session the *dominant theme* by looking at the theme of all the posters in a category and checking which theme has the largest number of posters. The dominant themes are used to color the category markers on the displays. From the assignments of the 12 844 posters, lists of 7 themes, 71 subthemes, 415 topics and 650 sessions were built. We are primarily interested in the visualization of the above categories in order to provide an overview of the field and check whether the ensuing groupings of posters into categories are homogeneous and naturally cluster in the Term Spaces defined in the following Section 2.1. Two visualization techniques were used: 3D-SE viewer and multidimensional scaling, so that the particular advantages of each approach could be exploited.

2.1. The construction of Term Spaces

The *Vector Space Model* (Salton, Wong, & Yang, 1975) is a common approach in Natural Language Processing, it is used to model textual documents. In this model, a set of terms $\mathcal{T}$ is first built by extracting all words occurring in a collection of documents $\mathcal{D}$, followed by stop words removal and stemming steps (Porter, 1980). The number of occurrences of each term in each document (usually called *frequency*) is counted and denoted f_{ij}. Then a frequency matrix **F** is built with the $\{f_{ij}\}$ in entries, as a [*terms* $\times$ *documents*] matrix or as a [*documents* $\times$ *terms*] matrix, where each document is a row vector in the space of all terms occurring in documents. This space of all terms is called *Term Space* in the present paper. Depending on the size of the Term Space, terms occurring too often or very seldom in documents can be discarded. When the number of documents N in the collection is in the range of a few thousands, the number of extracted terms M is often in the range of tens of thousands, leading to very high dimensional Term Spaces. In order to reduce the Term Space dimensionality, it is necessary to remove less semantically significant terms by keeping only a subset of the extracted terms, which was done using a ranking of the terms according to their Document Frequency scores (denoted DF hereafter). In general, we are interested in selecting the terms that best represent the semantic content of the documents. This intuitive feature is however very difficult to catch only by means of statistics. Two different sources of information from which words were extracted to build the Term Spaces are presented here below. Generated Term Spaces, identified hereafter by their dimension M, and the basic features of the corresponding data matrices are summarized in Table 2.

2.1.1. Terms extracted from the posters' abstracts and titles

The posters abstracts and titles were extracted from a CD-ROM distributed to all the participants of the Annual Meeting. The terms extracted from the posters' titles and abstracts were considered as equally important, that is we did not apply any extra weighting of title terms (e.g. by doubling the occurrence frequencies of title terms), as used sometimes to reflect a greater semantic importance of titles. Using the same preprocessing scheme and extraction of candidate terms as in Usui, Palmes, Nagata, Taniguchi, and Ueda (2007), a number $M = 40\,767$ of terms were extracted directly from the abstracts and titles of the $N = 12\,844$ posters. The number of terms in each document varies from 61 to 456, with an average of 278.86 terms per document. This space is much too large to allow further processing. A smaller Term Space was built by selecting terms occurring in at least 45 documents ($DF \geq 45$), in order to reduce the Term Space size to $M = 3006$ terms. For the sake of simplicity, only unigrams (single words) were considered as terms in this study.