Information Fusion for Scientific Literature Classification

Gary G. Yen
Oklahoma State University, USA

INTRODUCTION

Scientific literatures can be organized to serve as a roadmap for researchers by pointing where and when the scientific community has been and is heading to. They present historic and current state-of-the-art knowledge in the interesting areas of study. They also document valuable information including author lists, affiliated institutions, citation information, keywords, etc., which can be used to extract further information that will assist in analyzing their content and relationship with one another. However, their tremendously growing size and the increasing diversity of research fields have become a major concern, especially for organization, analysis, and exploration of such documents. This chapter proposes an automatic scientific literature classification method (ASLCM) that makes use of different information extracted from the literatures to organize and present them in a structured manner. In the proposed ASLCM, multiple similarity information is extracted from all available sources and fused to give an optimized and more meaningful classification through using a genetic algorithm. The final result is used to identify the different research disciplines within the collection, their emergence and termination, major collaborators, centers of excellence, their influence, and the flow of information among the multidisciplinary research areas.

BACKGROUND

In addition to the body content, which is sometimes hard to analyze using a computer, scientific literatures incorporate essential information such as title, abstract, author, references and keywords that can be exploited to assist in the analysis and organization of a large collection (Singh, Mittal & Ahmad, 2007; Guo, 2007). This kind of analysis and organization proves helpful while dealing with a large collection of articles with a goal of attaining efficient presentation, visualization, and exploration in order to search for hidden information and useful connections lying within the collection. It can also serve as a historic roadmap that can be used to sketch the flow of information during the past and as a tool for forecasting possible emerging technologies. The ASLCM proposed in this study makes use of the above-mentioned types of information, which are available in most scientific literatures, to achieve an efficient classification and presentation of a large collection.

Many digital libraries and search engines make use of title, author, keyword, or citation information for indexing and cataloging purposes. Word-hit-based cataloging and retrieval using such types of information tends to miss related literatures that does not have the specified phrase or keyword, thus requiring the user to try several different queries to obtain the desired search result. In this chapter, these different information sources are fused to give an optimized and all-rounded view of a particular literature collection so that related literatures can be grouped and identified easily.

Title, abstract, author, keyword, and reference list are among the most common elements that are documented in typical scientific literature, such as a journal article. These sources of information can be used to characterize or represent literature in a unique and meaningful way, while performing computation for different information retrievals including search, cataloguing or organization. However, most of the methods that have been developed (Lawrence, Giles & Bollacker, 1999; Morris & Yen, 2004; White & McCain, 1998; Berry, Dramac & Jessup, 1999) use only one of these while performing the different information retrieval tasks, such as search and classification, producing results that focus only on a particular aspect of the collection. For example, usage of reference or citation information leads to a good understanding of the flow of information within the literature collection. This is because most literatures provide a link to the original base knowledge they used

within their reference list. In a similar fashion, use of information extracted from the authors list can lead to a good understanding of various author collaboration groups within the community along with their areas of expertise. This concept can be extended analogously to different information types provided by scientific literatures.

The theory behind the proposed ASLCM can be summarized and stated as follows:

- Information extracted from a particular field (e.g., citation alone) about a scientific literature collection conveys limited aspect of the real scenario.
- Most of the information documented in scientific literature can be regarded useful in some aspect toward revealing valuable information about the collection.
- Different information extracted from available sources can be fused to infer a generalized and complete knowledge about the entire collection.
- There lies an optimal proportion in which each source of information can be used in order to best represent the collection.

The information extracted from the above mentioned types of sources can be represented in the form of a matrix by using the vector space model (Salton, 1989). This model represents a document as a multi-dimensional vector. A set of selected representative features, such as keywords or citations, serves as a dimension and their frequency of occurrence in each article of interest is taken as the magnitude of that particular dimension, as shown in Equation (1).

$$M = \begin{array}{c} \\ D_1 \\ D_2 \\ \dots \\ D_n \end{array} \begin{array}{c} \begin{array}{cccc} T_1 & T_2 & \dots & T_t \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1t} \\ a_{21} & a_{22} & \dots & a_{21} \\ \dots & & & \\ a_{n1} & a_{n2} & \dots & a_{nt} \end{bmatrix} \end{array}. \quad (1)$$

In the above matrix, hereafter referred to as *adjacency matrix*, every row corresponds to a particular document and every column to a selected feature. For example, if t terms are chosen as representative features for n documents, $a_{i,j}$ corresponds to the frequency of occurrence of term j in document i. This technique of document modeling can be used to summarize and represent a collection of scientific literatures in terms of several adjacency matrices in an efficient manner. These adjacency matrices are later transformed into similarity matrices that measure the inter-document similarities so that classification, search, or other information retrieval tasks can be performed efficiently. The procedure of similarity matrix computation can be carried out by using either cosine coefficient, inner product, or dice coefficient (Jain, Murty & Flynn, 1999).

Selection of the type of similarity computation method to be used depends on the nature of the feature and the desired purpose. Once the different available inter-document similarity matrices are calculated, the information contained in each matrix can be fused into one generalized similarity matrix that can be used to classify and give a composite picture of the entire collection.

MAIN FOCUS

This chapter presents an information fusion scheme at the *similarity matrix* level in order to incorporate as much information as possible about the literature collection that would help better classify and discover hidden and useful knowledge. Information fusion is the concept of combining information obtained from multiple sources such as databases, sensors, human collected data, etc. in order to obtain a more precise and complete knowledge of a specific subject under study. The idea of information fusion is widely used in different areas such as image recognition, sensor fusion, information retrieval, etc. (Luo, Yih & Su, 2002).

Similarity Information Gathering

The scope of this research is mainly focused on similarity information extracted from bibliographic citations, author information, and word content analysis.

Bibliographic Citation Similarity: Given a collection of n documents and m references, an $n \times m$ paper-reference representation matrix PR can be formed, where P stands for paper and R for references. Here usually m tends to be much larger than n because a paper commonly cites more than one reference and different papers have different reference lists. An element of the PR matrix, $PR(i, j)$, is set to one if reference j is cited in paper i. As a result, this matrix is normally a sparse matrix with most of its entities having a value of zero. Having this PR matrix, the *citation similarity*