Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 42 (2006) 4–30
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

A general matrix framework for modelling Information Retrieval

Thomas Rölleke *, Theodora Tsikrika, Gabriella Kazai

Department of Computer Science, Queen Mary University of London, London E1 4NS, UK

Accepted 12 November 2004
Available online 5 January 2005

Abstract

In this paper, we present a well-defined general matrix framework for modelling Information Retrieval (IR). In this framework, collections, documents and queries correspond to matrix spaces. Retrieval aspects, such as content, structure and semantics, are expressed by matrices defined in these spaces and by matrix operations applied on them. The dualities of these spaces are identified through the application of frequency-based operations on the proposed matrices and through the investigation of the meaning of their eigenvectors. This allows term weighting concepts used for content-based retrieval, such as term frequency and inverse document frequency, to translate directly to concepts for structure-based retrieval. In addition, concepts such as pagerank, authorities and hubs, determined by exploiting the structural relationships between linked documents, can be defined with respect to the semantic relationships between terms. Moreover, this mathematical framework can be used to express classical and alternative evaluation measures, involving, for instance, the structure of documents, and to further explain and relate IR models and theory. The high level of reusability and abstraction of the framework leads to a logical layer for IR that makes system design and construction significantly more efficient, and thus, better and increasingly personalised systems can be built at lower costs.

Keywords: Information Retrieval; Content; Structure; Semantics; Matrix spaces; Frequency-based operations; tf-idf; Evaluation measures; IR models; Eigenvectors

1. Introduction

With the Web and its search engines, ranking of retrieved objects becomes a focus in many application areas. More and more people face the task of building complex information systems that provide ranking

* Corresponding author.
E-mail addresses: thor@dcs.qmul.ac.uk (T. Rölleke), theodora@dcs.qmul.ac.uk (T. Tsikrika), gabs@dcs.qmul.ac.uk (G. Kazai).

0306-4573/$ - see front matter © 2004 Published by Elsevier Ltd.
doi:10.1016/j.ipm.2004.11.006

functionality. In this paper, we present a matrix framework in which key Information Retrieval (IR) concepts (Baeza-Yates & Ribeiro-Neto, 1999; Belew, 2000; Grossman & Frieder, 1998; van Rijsbergen, 1979) are described. This matrix framework supports the construction of efficient, flexible and robust search systems, since the matrix operations provide a high level of reusability and abstraction. For a search system engineer, this flexibility of retrieval and indexing functions is crucial, since it yields the possibility to tune the effectiveness and efficiency of a system for the particular personalised needs of end users.

The major theoretical foundations and motivations for this framework include the generalised vector-space model (Wong & Yao, 1995) and the probabilistic framework (Wong & Yao, 1995) for IR. Furthermore, research on the duality of document indexing and relevance feedback (Amati & van Rijsbergen, 1998), on term frequencies normalisation (Amati & van Rijsbergen, 2002) and on link analysis ranking algorithms for Web IR (Kleinberg, 1999; Page, Brin, Motwani, & Winograd, 1998) motivated the development of our matrix framework. While these works, however, address the formalisation of either content or structure, we propose a general matrix framework for both content and structure together with semantics. In addition, we include the modelling of evaluation measures and of retrieval models.

Throughout the paper, particular emphasis is given to a well-defined notation of matrix norms and operations. This allows for the dualities of the matrices and spaces defined within the framework to be systematically explored. For instance, widely used frequencies, such as tf-idf term weighting used for content-based retrieval, can be applied on the structure of collections or documents. On the other hand, concepts, such as pagerank, authorities and hubs, determined by the relationships between the documents of a collection (the collection structure), can be transferred to the relationships between the terms of a collection (the collection semantics).

The paper is structured as follows. Section 2 introduces the content-based, structure-based and semantic-based aspects of retrieval. We consider collection, document, and query matrix spaces and define matrices and operations on them to express these retrieval aspects. Section 3 proposes frequency-based operations for expressing basic content-based IR concepts, such as term frequency and inverse document frequency. Section 4 shows how to use the general matrix framework for modelling classical and alternative evaluation measures. In Section 5, the general matrix framework is used for the modelling of retrieval models. Finally, Section 6 examines the meaning of the eigenvectors of the symmetric matrices and shows the dualities between collection, document and query spaces.

2. Retrieval aspects expressed in the matrix spaces

The underlying framework of our general IR model consists of matrix spaces, matrices associated with each element of a space and standard linear algebra operations on matrices. We consider three matrix spaces: a collection space, a document space and a query space. Each space may contain several elements. For example, the collection space contains collections and the document space contains documents. Each space has two dimensions. For example, the collection space has document and term dimensions, each represented by a vector. For each element of a space, we introduce matrices to represent the relationships between pairs of elements of the two dimensions of the space and matrices to represent the parent–child relationships [1] between pairs of elements of a single dimension of the space.

We propose a carefully chosen notation for indicating the spaces and their associated matrices. In our notation, a space is represented by a lower case letter and its dimensions by capital case letters. Let us consider a space s and its dimensions X and Y. The vectors X_s, Y_s contain the elements of the dimensions. The

[1] The terminology *parent–child relationship* is used to describe any directed association between a source (*parent*) and a target (*child*), i.e. it is not restricted to (tree-like) hierarchical relationships.