The Journal of Systems and Software 84 (2011) 2303–2313

Contents lists available at ScienceDirect

The Journal of Systems and Software

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

# Tensor Field Model for higher-order information retrieval

Qiao Ya-nan*, Qi Yong, Hou Di

*School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China*

ARTICLE INFO

*Article history:*
Received 27 March 2010
Received in revised form 21 June 2011
Accepted 21 June 2011
Available online 29 June 2011

*Keywords:*
Information retrieval
Text representation
Text mining

ABSTRACT

There is an implied assumption for keywords in the traditional Information Retrieval (IR) models: keywords are parallel to each other. In fact, there are some relations between terms in quite a few queries, and in these queries, perhaps one term is subordinate to another term according to the inner meanings of information needs. This is "Higher-order IR"(HIR) defined in this paper, and we call traditional IR "first-order IR" instead. Some research fields such as Public Opinion Analysis, Chain of Events Analysis and Trend Analysis which reflect the vague concept of HIR are all special form of HIR. Apparently, traditional IR models cannot deal with HIR directly. We need a new HIR model to represent and organize the documents, queries and relevance relationship between them. In this paper, we propose "Tensor Field Model" (TFM), and its perspectives are field theory in physics and multilinear algebra in maths. We construct the tensor representations of documents and queries in TFM, presenting some key concepts such as term field, tensor product of term array and term field constant. Empirical results show that TFM is appropriate for HIR theoretically and formally compared with traditional models which simplify the HIR problems as first-order IR problems to some extent.



## 1. Introduction

As core functions of information retrieval (IR), the representation and organization of the information items should provide the user with easy access to the information in which he is interested (Baeza-Yates and Ribeiro-Neto, 1999). Well then how to represent the users' information needs? Generally speaking, users must convert their information needs into a series of query terms which can be processed by IR system. For four classic IR models: Boolean Model, Vector Space Model (Salton et al., 1975), Probabilistic Model (Fuhr, 1992), Language Model (Croft, 2003) and their variants, the logical views for the users' information needs are all based on keywords mainly, and their key difference is the perspective on IR: perspective of set theory, linear algebra or probabilistic, etc.

In fact, there is an implied assumption in the traditional IR models: the terms in the query are parallel to each other. For example, user wants documents about the reactions of penicillin, so he perhaps inputs this query: "penicillin AND reaction". It works seemingly, and search engine will return some expected documents partly. However, these two query terms are not parallel obviously. User needs the documents in which terms "reaction" is subordinate to "penicillin". If terms "penicillin" and "reaction" are in the same document but they are two separate terms without grammatical relations, and then the document does not meet user's actual needs. In this paper, we call original IR problems without subordinations *first-order IR problems*, and call IR problems with first-order subordinations *second-order IR problems*.

There are more complicated situations. For instance, user wants documents which opposed impeaching Clinton in some affair, so he inputs the query: "Clinton AND impeach AND oppose". These three query terms are in second-order subordination clearly: "oppose" is subordinate to "impeach" and "impeach" is subordinate to "Clinton". Similarly, we call IR problems with second-order subordinations *third-order IR problems*. Generally, we call the IR problems with whose order are greater than or equal to 2 *Higher-order IR problems*.

IR problems with different orders can be compounded. For example, the user's information need is the documents about the reactions of penicillin and opposing abuse of antibiotics. This is a compound IR problem with second-order and third-order. The order of compound IR problems is the highest order of the subproblems.

Then we give three examples of Higher-order IR problems in some professional research fields:

1. *Public Opinion Analysis*: Public opinion is the aggregate of individual attitudes or beliefs held by the adult population. Governments have increasingly found surveying it to be useful tools for guiding their public information and propaganda programs and occasionally for helping in the formulation of other kinds of policies. Public Opinion Analysis is second-order IR problems

* Corresponding author. Tel.: +86 29 82664160.
*E-mail addresses:* qiaoyanan@mail.xjtu.edu.cn (Y.-n. Qiao), qiy@mail.xjtu.edu.cn (Q. Yong), houdi@mail.xjtu.edu.cn (H. Di).

0164-1212/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jss.2011.06.057

indeed: government expects to know public opinion of policies or affairs from some information channels (for example, Internet documents) in which "opinion"(a group of tendentious words, such as "oppose" or "support") is subordinate to "policies" or "affair".
2. *Chain of Events Analysis*: A chain of events is a number of actions and their effects that are contiguous and linked together by causality. There are three elements in chain of events: subject, actions of the subject and the time of the actions. If we want to extract a chain of events from Internet documents, we expect to retrieve documents in which the "time" is subordinate to "action" and "action" is subordinate to "subject". So Chain of Events Analysis is third-order IR problems.
3. *Trend Analysis*: It refers to the concept of collecting information and attempting to spot a pattern or trend in the information. Trend Analysis is also a second-order IR: such documents are expected, in which "trend" ("increase", "weaken", etc.) is subordinate to "subject".

These examples above can be generalized to Higher-order IR problems because of some common features. Differing from traditional IR or first-order IR, the focus of Higher-order IR is not logical relations such as "AND", "OR" or "NOT", but the cascade relations of query terms. Some complex information needs as mentioned above must be denoted by query terms not only with logical relations but also with cascade relations.

Apparently, traditional IR models cannot deal with Higher-order IR directly. We need a new Higher-order IR model to represent and organize the documents, queries and relationships between them. In this paper, we propose *Tensor Field Model* (TFM). Its perspective is multilinear algebra, introducing some core concepts such as *term field*, *Tensor Product of Tensor Array* and *term field constant*. Empirical results show that TFM is appropriate to Higher-order IR theoretically and formally, compared with traditional IR models which simplify the problems to some extent.

With the new model for Higher-order IR, we get the following preliminary contributions:

1. The proposed new concept of Higher-order IR. It enables us to analyze the cascade relations of the terms, accessing implied information in the documents. From three examples of Higher-order IR mentioned above we can conclude that the vague concept of Higher-order IR has been formed for a long time but lack of unified analyses and discussions.
2. The physical interpretations for the query terms relations with open definitions. It is different to GBM (Shi et al., 2005) which also did it but was limited to "Gravitation".
3. The Tensor Field Model as an initial solution for Higher-order IR.

The rest of the paper is organized as follows. We first introduce the main work of related researchers in Section 2. In Section, we consult some basic concepts and propose Tensor Field Model. Then we test these hypotheses with systematic experiments in Section 4. Finally, give conclusions and future research directions in Section 5.

## 2. Related work

In this section we will outline two kinds of related work: some are research about cascade relations of query terms; others are two *Tensor Space Model*, which generalized VSM from another point of view, so we called the model proposed in this paper *Tensor Field Model* to show differences.

### 2.1. Related work about cascade relations

Shi proposed *Gravitation-Based Model* (GBM), a physical model for information retrieval inspired by Newton's theory of gravitation (Shi et al., 2005). In GBM, documents and queries are modeled as objects and the relationship between a query and a document is modeled as the attractive force between them.

Metzler developed a formal framework for modeling term dependencies via Markov random fields (Metzler and Croft, 2005). In his work, three variants of term dependencies model are described: full independence (query terms are independent), sequential dependence (dependence between neighboring query terms) and full dependence (all query terms are dependent on each other in some way), and then use the linear combination of these variants to rank documents.

Beigbeder proposed a model based on a fuzzy proximity degree of term occurrences (Beigbeder and Mercier, 2005). He defined a concept of proximity between a position in the document and a term, which is close to the concept of term field in Section 3.1.1.

Petkova's work is a expert search model based on the dependency between the named entities and terms which appear in document (Petkova and Croft, 2007), and it is a second-order IR problems for documents in which "specific topic" is subordinate to "expert". She proposed several proximity kernels to form the dependency between the named entities and terms, such as triangle kernel, Gaussian kernel and step function.

Rasolofo and Savoy (2003) proposed a combination of proximity measurement and Okapi probabilistic model (Robertson and Walker, 1994). His approach copes with multi-term queries and assumes the probability that a document will be relevant must be greater if the document contains sentences having at least two query terms within them.

Chen proposed a new keyword suggestion method that fully exploits the semantic knowledge among concept hierarchy (Chen et al., 2008). He matched a given keyword with some relevant concepts firstly, and then used these relevant concepts with their hierarchy to mine the inner information of the keywords. Finally new keywords are suggested according to the inner information. In Chen's work, the hierarchy relation is an inherent property of concepts, and it is different from other related work in this section. In these work, cascade relation (or proximity relation) is a temporary status of two terms in a specific language environment.

### 2.2. Related work about "Tensor Space Model"

Aiming at a keyword "vector" in the name of "Vector Space Model", some researchers proposed *Tensor Space Model* based on the generalization of "vector", namely "tensor", to improve traditional Vector Space Model.

Liu et al. (2005) regarded all characters except 26 English letters as blanks, and all 27 characters, including 26 English letters and blank, are regarded as coordinates in space. All text in English can be regarded as characters stream, and strings consist of every adjacent $N$ characters can be converted to a series of points in $N$-dimension space, so whole text can be converted to a $N$-order tensor with $27^N$ elements. A corpus ($m$ documents) can be converted to a $(N+1)$-order tensor with $m \times 27^N$ elements, and then analyze them using similar methods in Vector Space Model.

Cai et al. (2006) noticed that when the documents in corpus are more and more, the dimension of vector space model is bigger and bigger, so do the complexity of algorithm. So he converted a $N$-dimension document vector to a $N_1 \times N_2$ matrix, $N \approx N_1 \times N_2$, regarding it as the second-order tensor representation of the document, and then proposed a new algorithm "TensorLSI".

Liu and Cai proposed two different "Tensor Space Model" respectively with the same name, which are novel ideas and achieved