Information Systems 50 (2015) 20–35

Contents lists available at ScienceDirect

Information Systems

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

A geometric framework for data fusion in information retrieval

Shengli Wu [a,b,*], Fabio Crestani [c]

[a] School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, China
[b] School of Computing and Mathematics, University of Ulster, Newtownabbey, UK
[c] Faculty of Informatics, University of Lugano, Switzerland

ARTICLE INFO

Article history:
Received 15 July 2010
Received in revised form
28 August 2013
Accepted 5 January 2015
Available online 12 January 2015

Keywords:
Database searching
Geometric modeling
Information retrieval
Data fusion

ABSTRACT

Data fusion in information retrieval has been investigated by many researchers and a number of data fusion methods have been proposed. However, problems such as why data fusion can increase effectiveness and favorable conditions for the use of data fusion methods are poorly resolved at best. In this paper, we formally describe data fusion under a geometric framework, in which each component result returned from an information retrieval system for a given query is represented as a point in a multi-dimensional space. The Euclidean distance is the measure by which the effectiveness and similarity of search results are judged. This allows us to explain all component results and fused results using geometrical principles. In such a framework, score-based data fusion becomes a deterministic problem. Several interesting features of the centroid-based data fusion method and the linear combination method are discussed. Nevertheless, in retrieval evaluation, ranking-based measures are the most popular. Therefore, this paper investigates the relation and correlation between the Euclidean distance and several typical ranking-based measures. We indeed find that a very strong correlation exists between these. It means that the theorems and observations obtained using the Euclidean distance remain valid when ranking-based measures are used. The proposed framework enables us to have a better understanding of score-based data fusion and use score-based data fusion methods more precisely and effectively in various ways.

1. Introduction

In information retrieval, numerous retrieval models have been developed and a variety of representation methods for queries and documents proposed to enhance retrieval effectiveness. These techniques are often comparable in performance and there is no overall winner. This naturally leads to the question: how can we maximize the use of resources at our disposal? Well, we may use a number of independent information retrieval systems or use one single retrieval system but several different query representations or parameter settings to retrieve a given document collection for an arbitrary query. Then we can merge all the results for better retrieval effectiveness. This is the primary idea behind data fusion.

Up to now, data fusion has been widely used in many different IR-related tasks, such as developing new retrieval models [34], web searches [20], retrieval evaluation [33], text summarization [48], and others.

Considerable effort has gone into investigating data fusion in the information retrieval community. Most previous research empirically investigates the topic. Key questions such as why data fusion can improve retrieval effectiveness and the optimal conditions for data fusion are only partially or vaguely answered. For example, a

* Corresponding author.
E-mail addresses: swu@ujs.edu.cn, s.wu1@ulster.ac.uk (S. Wu), fabio.crestani@unisi.ch (F. Crestani).

http://dx.doi.org/10.1016/j.is.2015.01.001
0306-4379/

plethora of observations and hypotheses (such as Lee's hypothesis [22], the skimming effect, the chorus effect, and the dark horse effect mentioned by Vogt and Cottrell [46,47], the authority effect and the rank effect mentioned by Spoerri [42,43]) have been made or put forward, yet none of them hold the whole time. Another problem with these observations and hypotheses is that they are "ambiguously" stated. This means that we do not know how best to exploit them. Moreover, previous research [31,46,47,50,51] finds that higher similarity among results detract from the efficacy of data fusion. However, such observations are based on statistically large sets of data and thus do not necessarily hold in all individual fusion cases. Although a finding like this may have its uses in data fusion (e.g., [51]), we know not how best to use it.

We find that a major reason for this problem is the inadequacy of current measures used for retrieval evaluation. At present, almost all commonly used measures such as average precision, recall-level precision, etc., are ranking-based measures. Those ranking-based measures are only concerned with the relative positions of relevant/irrelevant documents, and make it difficult for us to have a thorough understanding of the nature of data fusion in information retrieval.

In this paper, we set up a theoretical framework for data fusion in information retrieval, in which all results (either from information retrieval systems or from data fusion methods) are represented as points in a multi-dimensional space. This immediately presents us with an added nicety: the similarity between results and the effectiveness of any result can be evaluated by the same measure – the Euclidean distance. Consequently, data fusion in information retrieval becomes a deterministic problem and we can prove many useful results.

The rest of this paper is organized as follows. Section 2 offers a review of some related work, while Section 3 discusses the uncertainty of the effectiveness of the fused results when ranking based measures are used. Section 4 introduces the framework of data fusion based on geometrical principles. In Sections 5 and 6 we further discuss some characteristics of the centroid-based data fusion method and the linear combination method under this framework. Section 7 discusses the relationship between the Euclidean distance and ranking-based measures. Section 8 is the conclusion.

2. Related work on data fusion

Some early related work on data fusion was from Saracevic and Kantor [40], Turtle and Croft [45], Foltz and Dumais [11], Bartell et al. [3], and Belkin and his colleagues [5,6]. Their experiments demonstrated that data fusion improved performance in a range of different settings.

Data fusion methods can be divided into two categories: relevance score-based methods and ranking-based methods. The division mainly depends on the information type required from component retrieval systems.

A fair number of possibilities have been looked into. Aslam and Montague studied Borda count in [2], Markov chain-based methods were investigated by Dwork et al. in [9] and Renda and Straccia in [35], Montague and Aslam examined Condorcet fusion in [29], a probabilistic approach was considered by Lillis et al. in [24], Farah and Vanderpooten proposed an outranking approach that mixed multiple hypotheses in [10], whilst Wu et al. explored a cubic regression model-based approach in [53]. Among these, [2], [24] and [53] used various methods to assign scores to documents at different ranks, and then used relevance score-based methods to fuse results. Hsu and Taksa [16] compared the performance of ranking-based methods and relevance score-based methods.

In the following we review some more work on data fusion that all use relevance scores. Fox and co-workers [12,13] introduced a group of data fusion methods including CombSum and CombMNZ. CombSum sets the score of each document in the combination to the sum of the scores obtained by the component results, while in CombMNZ the score of each document is obtained by the product of this sum and the number of results that have nonzero scores.

Lee [22] conducted an experiment with 6 runs selected from TREC 3 to support his hypothesis: different retrieval processes might retrieve similar sets of relevant documents but retrieve different sets of non-relevant documents. Furthermore, Lee stated that an improvement in performance could be attained as long as the component results being used for fusion had greater relevant overlap than non-relevant overlap. He used this hypothesis to explain why CombMNZ was an effective data fusion method.

Beitzel et al. [4] therefore set out to evaluate the performance of CombMNZ using several different groups of systems. They observed no improvement when fusing results from three different retrieval strategies in the same information retrieval system, while the merged result was better than the best component system when choosing the top three systems submitted to TREC 6, 7, 8, 9 and 2001, though the condition that greater relevant overlap than non-relevant overlap in component results was always satisfied. Their experimental results disproved Lee's hypothesis.

Vogt and Cottrell [46,47] analyzed the performance of the linear combination method, which combines multiple component results by linear regression. In their experiments, they used all possible pairs of 61 systems submitted to the TREC 5 ad-hoc track. The similarity of two results' rankings and 13 other variables were used in the analysis. They noted three interesting phenomena:

- The skimming effect happens when different retrieval systems retrieve different relevant documents, so that a fusion method that takes the top-ranked documents from each of the retrieval systems will force non-relevant ones down in the rankings.
- The chorus effect occurs when several retrieval systems suggest that a document is relevant to a query, this tends to be stronger evidence for relevance than a single system doing so.
- The dark horse effect means that a retrieval system may produce unusually accurate (or inaccurate) estimates of relevance for at least some documents, relative to the other retrieval systems.

They reasoned that a good data fusion method should be able to exploit all these effects. However, they observed