情报工程 ISSN 2095-915X
第2卷 第3期 37-42 2016 年 6 月
TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN 2095-915X Vol.2 No.3 37-42 Jun 2016

doi:10.3772/j.issn.2095-915x.2016.03.005

基于 VSM 的科技期刊文献与专利文献的相似度计算方法研究

曾文，徐红姣，李颖，王莉军，赵婧
（中国科学技术信息研究所 北京 100038）

摘要：文本相似度的计算方法以采用 TF-IDF 的方法对文本建模成词频向量空间模型（VSM）为主，本文结合科技期刊文献和专利文献特点，对 TF-IDF 的计算方法进行了改进，将词频的统计改进为科技术语的频率统计，提出了一种针对科技文献相似度的计算方法，该方法首先应用自然语言处理技术对科技文献进行预处理，采用科技术语的自动抽取方法进行科技文献术语的自动抽取，结合该文提出的术语权重计算公式构建向量空间模型，来计算科技期刊文献和专利文献之间的相似度。并利用真实有效的科学期刊和文献数据进行实验测试，实验结果表明文中提出的方法优于传统的 TF-IDF 计算方法。

关键词：自然语言处理，TF-IDF，向量空间模型，科技期刊，专利，相似度

中图法分类号：TP311

The Study of Correlation Calculation Method Based on the VSM for Scientific and Technological Periodicals and Patents

ZENG Wen, Xu HongJiao, Li Ying, Wang LiJun, Zhao Jing
(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: Original text similarity measurements employed the TF-IDF method to model the documents as term frequency vector space model (VSM), and compute similarity between the documents. The paper proposed a

基金项目：本研究得到国家社会科学基金项目（项目编号：14BTQ038）和中国科学技术信息研究所科研项目预研资金项目（项目编号：YY2016-08）的支持。
作者简介：曾文（1973-），博士，副研究员，硕士生导师，研究方向：知识组织，智能信息处理；徐红姣（1985-），硕士，助理研究员，研究方向：信息关联分析与检索技术；李颖（1965-），博士，副研究员，硕士生导师，研究方向：知识组织，数字图书馆；王莉军（1978-），博士，助理研究员，研究方向：数据挖掘；赵婧（1987-），硕士，助理研究员，研究方向：情报学。

new literature similarity calculation method for scientific and technological (S&T) documents. According to the characteristics of these documents, we replaced the word frequency statistic method by the scientific term frequency statistic method to improve the algorithm of TF-IDF method. In addition, the new method applied the natural language processing technology to the pretreatment, using the term automatic extraction method for extracting S&T terms. The term weight VSM was constructed to calculate the similarity between S&T periodical literatures and patents by using the new calculation formula. Moreover, this paper used the real S&T documents to test the new method, and compared its results with the original method. The results showed that the proposed method is superior to the original TF-IDF method.

Keywords: Natural language processing, TF-IDF, vector space model, journal of science and technology, patent, similarity

1 引言

国家和政府对科技文献数据资源的建设非常重视，2012 年 1 月 30 日刘延东同志视察中国科学技术信息研究所时强调，加快科技信息事业发展，既是应对日益激烈的国际竞争、实现创新驱动的迫切要求，也是提高我国科技改革发展的基础水平和创新能力的客观需要。如何帮助用户全面、快速、准确地发现蕴含在文献中的情报（如技术和知识），从不同的情报维度展示这些文献所承载的情报之间的各种关联，进而辅助用户更高效地进行技术创新，这是实现下一代科技文献服务跨越式发展的必要条件之一。在大数据环境下，科技数据来源分布广泛、数据质量良莠不齐，数据内容深度千差万别。科技大数据呈现的特点是：数据量大且增长速度快；数据来源和数据结构类型多；有价值的数据比例小；数据具有敏感性和积累性，会涉及国家安全和利益；数据管理和数据分析具有复杂性。而科技文献的价值评价的基础问题之一是建立相同类型和不同类型科技文献之间的关联关系，分析学术和应用之间的关系，挖掘这种关联关系的技术之一是文献之间相似度的计算问题。我们知道：科技期刊文献和专利文献是科技文献中最重要的两类文献。本文认为：对科技期刊文献和专利文献的相似度研究对于评价科技期刊文献和专利文献的学术价值和应用价值具有重要的参考价值，且对于目前单纯依赖于外在指标运用文献计量学方法进行文献的评价，将是一个有益的补充。基于此，本文重点以科技期刊和专利之间的相似度计算方法为例进行了相关研究。

2 研究现状

目前国内外对于科技文献之间，特别是不同类型科技文献之间的相似度计算研究并不多见，而以文本相似度计算问题的研究居多。尽管科技文献在结构和内容上与已有研究的文本是有区别的，但是科技文献之间相似度计算的方法与已有文本相似度的计算方法仍有异曲同工之处。文本的相似度计算是在语言学、心理学和信息管理等领域被广泛研究的问题之一。在信息检索领域、图像检索领域、文本摘要的自动生成、文本分类和文本的重复度检测等均具有广泛的应用。目前国内外关于文本相似度的研究以基于统计的方法应用最为广泛，常用的方法是向量空间模型

(VSM)，即将文本集中的每个文本表示成一个文本词语的特征向量，词语特征的权重是通过词频来计算的，通过向量的降维处理与相似度计算来衡量向量空间中文本特征的相似性。两个文本的词语之间的语义相似度为对应的特征间的相似度，两个文本的词语特征的相似度可以根据特征向量空间中两点间的距离进行度量。其中文本词语特征向量的距离计算最常用的方法为向量的余弦计算法。这种计算方法简单直观的反应文本直接的相似度，

但是它需要具体的文本数据语料集的支持，语料集的数据质量对统计计算结果的干扰性相对比较大。另外一种方法是运用“WordNet”，知网等知识模型，将本体论引入文本相似度的计算中，利用语义资源计算方法改善相似度计算的准确性，但因为需要构建本体，且本特本身受人为的主观影响较大，因此统计计算方法仍然是一种较好的文本相似度衡量方法。

在已有国内外的相关研究工作中，如 Shibata [3] 等人使用基于 TF−IDF 的文本相似度计算方法对太阳能领域的科技文献和专利分别进行了文献聚类分析和计算；黄承慧等 [4] 基于 TF−IDF 的文本相似度计算方法，金博 [5]、翟延冬 [6]、张佩云 [7] 等利用知网语义相似度的技术来实现文本之间的相似度计算方法。但是无论是基于 TF−IDF 的文本相似度计算方法，还是基于本体的语义概念计算方法，他们主要存在的问题是：一是这些研究工作均是基于相同类型的文本之间的相似度计算，二是文本内容不是以科技文献为对象开展的研究。例如，在以本体技术为基础的研究方法中，本体中的义原或词汇均不是科技领域内的专业术语和语义关系。所以本文认为完全沿用以有的文本相似度的计算方法是不可取的，应结合科技文献的特点进行科技文献之间的相似度计算。

3 基于向量空间模型的文本关联度计算方法

3.1 向量空间模型

向量空间模型 (VSM) 是一种代数模型。在近 30 年内，已被广泛应用到信息检索、文本分类、文本聚类等领域，并取得了较好的实验和应用效果。其基本思想是：用向量表示文本，每个维度对应于一个单独的词，且词与词之间是不相关的，文档 d_k $(w_1, w_2, w_3, \ldots, w_n)$ 可以表达成相互独立的词条 $(t_1, t_2, t_3, \ldots, t_n)$，为了表示词条的重要程度，为每个词条赋予相应的权值 w_i，其中文档 d_k 可用向量 $(w_1, w_2, w_3, \ldots, w_n)$ 表示。向量空间模型中的文档相似度计算方法为：

$$sim(d_k, d_p) = \frac{\vec{V}(d_k) \cdot \vec{V}(d_p)}{\left|\vec{V}(d_k)\right| \left|\vec{V}(d_p)\right|} = \frac{\sum_{i=1}^{n} w_{ki} \times w_{pi}}{\sqrt{\sum_{i=1}^{n} w_{ki}^2} \times \sqrt{\sum_{i=1}^{n} w_{pi}^2}} \quad (1)$$

3.2 基于向量空间模型的科技期刊与专利相似度计算

传统的文本相似度计算使用向量空间模型时，通常是将文本表示为文中出现的 N 个加权词组成的向量，其中每个词的权值 w 用词频（Term Frequency, TF），逆文本频率（Inverse Document Frequency, IDF）进行计算，公式如下：

$$TF - IDF(w_i) = tf(w_i) \times idf(w_i) = tf_i(w_i) \times \log(N / df(w_i)) \quad (2)$$

公式（1）中 $tf_i(w_i)$ 表示词 wi 在文本 j 中出现的频率，N 表示文本集合中所有文本的总数，$df(w_i)$ 表示词 wi 在文本集合中出现的总数。

在向量空间模型中，如果使用文本中各个字（词）的 TF−IDF 值来表示一个向量文本，并进行文本的相似度计算，则这个文本向量是高维而且极度稀疏的，本文认为：从每一篇科技文献

中抽取科技术语，以此来表示文献，既不影响文献特征提取，又尽可能地减少文献特征向量表示的维度，科技文献术语抽取的具体方法可参见文献 [7]。

基于向量空间模型的科技期刊与专利相似度计算的实现是通过文献之间相似度值的大小来体现的，在本文的方法中科技文献之间的相似度计算即由科技术语向量之间的相似度计算实现的，由于科技期刊文献与专利文献在结构和篇幅上存在差异，我们对向量的权重值的计算即公式（2）做了如下改进，详见公式（3）

其中，

df_{-t-m}：词 t 在科技期刊文献中出现的次数

df_{-t-N}：词 t 在专利文献中出现的次数

L_{d-avg}：所有科技期刊文献的平均长度

L_d：单篇科技期刊文献长度（t），即含词 t 的文献的长度

L_p：单篇专利文献长度（t）即含词 t 的专利文献的长度

L_{p-}：所有专利文献的平均长度

$$w_i = \sum_{t \in D,P} \left[\log \frac{M}{df_{-t-M}} \right] \cdot \frac{(k_1 + 1) df_{-t-m}}{k_1 \left((1 - b) + b \times \frac{L_d}{L_{d-avg}} \right) + tf_{dt}} \times \left[\log \frac{N}{df_{-t-N}} \right] \cdot \frac{(k_1' + 1) df_{-t-N}}{k_1' \left((1 - b') + b \times \frac{L_p}{L_{p-avg}} \right) + tf_{ps}} \quad (3)$$

4 基于向量空间模型的科技期刊与专利文献相似度计算的方法与实现

4.1 基于向量空间模型的科技期刊与专利文献相似度计算的基本处理流程

需要处理的科技期刊与专利文献原则上含有完备的文本信息，但是现有的自然语言处理技术不能完全处理这些文本信息，所以在使用向量空间模型表示文档之前，首先需对文献进行适当的预处理，主要包括中文分词、英文词根还原、去停用词和术语抽取，之后通过建立数据的索引，实现文献相似度的计算。具体实现流程如图 1 所示。

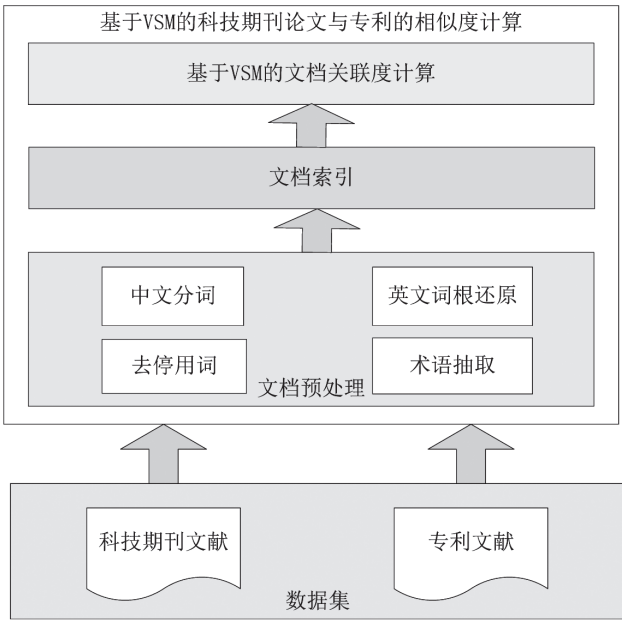

图 1　流程图

4.2 实验与分析

实验数据取自国家科技图书文献中心（NSTL）的真实科技期刊与专利文献数据源，我们以光学传感器（optical sensor）为检索词抽取 1989 年 至 2013 年 的中英文科技期刊和专利文献，经过清洗处理最终得到中文科技期刊文献 1604 篇，英文期刊文献 405 篇，中文专利 22 篇，英文专利 79 篇。图 2 和图 3 分别是我们抽取的中英文科技期刊与专利文献数据数量对比图示。

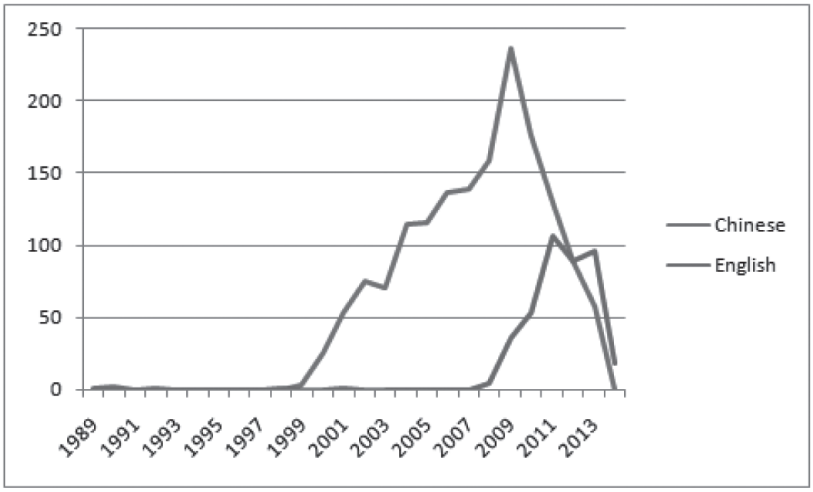

图 2 中英文科技期刊文献数量对比

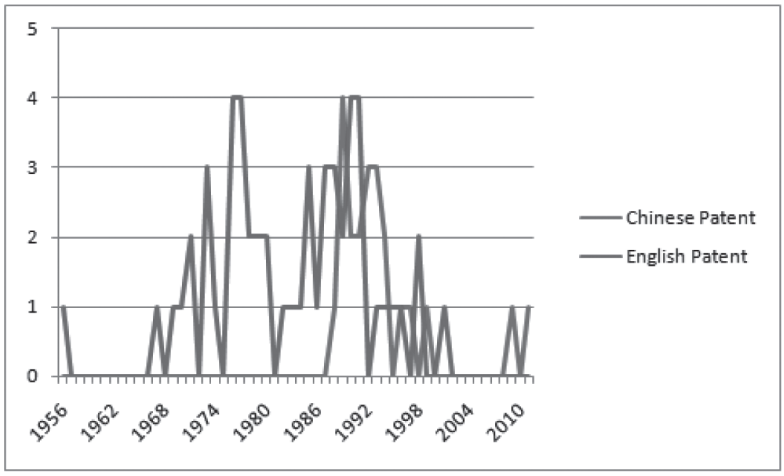

图 3 中英文专利文献数量对比

通过运用前文所述的计算方法，我们对科技期刊与专利文献数据进行了处理和相似度的计算，我们将每篇科技期刊文献与专利进行相似度的计算，相似度的值按高低进行排序，并对实验结果进行准确率的评价。准确率的计算公式如下：

$$Precision = \frac{CorrectMatch\_number}{Total\_number} \times 100\% \quad (4)$$

具体的实验结果比较如表 1 所示。

表 1 实验结果比较

名称	VSM（传统）	VSM（本文提出的）
中文期刊和专利	约 18.1%	约 31.8%
英文期刊和专利	约 10%	约 22.2%

5 结论

为了进行科技文献之间的关联分析，本文提出了一种针对科技期刊与专利文献进行相似度计算的方法，并以科技期刊文献和专利文献之间的相似度的计算为例，进行了实验研究。通过与传统的基于向量空间模型的文本相似度计算方法相比，本文做了两方面技术改进：一是表示科技文献文本信息的向量空间单元以科技术语为单位，而不是传统方法中的以词（字）为单位，二是对权重的计算方法进行了改进。实验结果表明对于科技文献相似度的计算，本文的提出的方法要优于传统的基于向量空间模型的方法。但是也应看到，由于考虑计算规模问题，我们的数据集选择在量的规模还不够，另一方面在语义的深度上还有待于进一步挖掘。未来我们将对研究的方法上做深层次的分析和研究工作，这对于提供科技文献信息服务将有着积极的应用意义。

参考文献

[1] 梅家驹 . 同义词词林 [M]. 上海 : 上海辞书出版社 , 1983.

[2] 董振东 , 董强 .《知网》[EB/OL]. [2016-05-04]. http://www.keena ge.com.

[3]Shibata N, Kajikawa Y, Sakata I. How to measure the semantic similarities between scientific papers and patents in order to discover uncommercialized research fronts: A case study of solar cells[C]// Technology Management for Global Economic Growth. IEEE, 2010:1-6.

[4] 金博 , 史彦军 , 滕弘飞 . 基于语义理解的文本相似度算法 [J]. 大连理工大学学报 , 2005, 45(2):291-297.

[5] 翟延冬 , 王康平 , 张东娜 , 等 . 一种基于 WordNet 的短文本语义相似性算法 [J]. 电子学报 , 2012, 40(3):617-620.

[6] 张佩云 , 陈传明 , 黄波 . 基于子树匹配的文本相似度算法 [J]. 模式识别与人工智能 , 2014, 27(3):226-234.

[7] 曾文 , 徐硕 , 张运良 , 等 . 科技文献术语的自动抽取技术研究与分析 [J]. 现代图书情报技术 , 2014(1):51-55.

[8]Schmoch U. Tracing the knowledge transfer from science to technology as reflected in patent indicators[J]. Scientometrics, 1993, 26(1):193-211.

[9]Callon M, Courtial J P, Laville F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry[J]. Scientometrics, 1991, 22(1):155-205.

[10]Magerman T, Looy B V, Song X. Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications[J]. Scientometrics, 2010, 82(2):289-306.