第 37 卷 第 15 期
Vol.37 No.15

计 算 机 工 程
Computer Engineering

2011 年 8 月
August 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)15—0134—03 文献标识码：A 中图分类号：TP391

基于场论的高精度信息检索研究

杨为民[a,b]，李龙澍[b]

(安徽大学 a. 智能计算与信号处理教育部重点实验室；b. 计算机科学与技术学院，合肥 230039)

摘 要：传统信息检索模型的信息获取准确率较低。针对该问题，提出一种基于场论的高精度信息检索模型，在考虑标引词之间相关性的前提下，以规范化标引词的数量作为文档信息量，构建一个文档信息场。该信息场以文档间相互作用的场力大小度量文档的相关性，由此得到更精确的检索结果。实验结果验证了该模型相对于同类模型的优越性。

关键词：信息检索；场论；高精度检索；文档相关性；检索模型

Research of High Accuracy Information Retrieval Based on Field Theory

YANG Wei-min[a,b], LI Long-shu[b]

(a. Key Lab of Intelligent Computing & Signal Processing of Ministry of Education;
b. School of Computer Science and Technology, Anhui University, Hefei 230039, China)

【Abstract】Traditional information retrieval model has the problem of low accuracy when obtaining the information. In order to solve the problem, this paper proposes a new information retrieval model based on field theory by analyzing the shortcomings of traditional information retrieval theory. By using the number of standardized index terms and taking account of the relation of index terms, a document information field is constructed. In the new field, the document relevance is computed by field force, so that the new model can get more accurate retrieval results than other models. Experimental results demonstrate its superiority.

【Key words】information retrieval; field theory; high accuracy retrieval; document relevance; retrieval model

DOI: 10.3969/j.issn.1000-3428.2011.15.042

1 概述

随着网络时代的到来，信息资源的数量呈现指数形式的增长。信息检索在信息获取中发挥了关键性的作用，但检索的准确性却有待进一步提高。自 2003 年的文本检索会议(Text Retrieval Conference, TREC)提出了高精度文档检索(High Accuracy Retrieval from Documents, HARD)后，国内外的学术界和企业界如中科院、马萨诸塞大学、约克大学、印第安纳大学、IBM、微软研究院等对此进行了大量的研究，分别从用户辅助查询的扩展、标引相关性分析、检索结果反馈、自组织映射的可视化解释等方面提高信息检索的精度[1]。本文从文档相关性的角度提出基于场论的信息检索模型，对反馈结果进行精度度量，从而动态提高检索结果的精度。

2 传统的信息检索模型

20 世纪 60 年代以来，信息检索领域在索引模型、文档内容表示、匹配策略等方面取得了许多重要的成果，提出了大量的检索模型[2]。从信息检索的数学基础分析，信息检索模型主要分为 3 类：集合论模型，代数模型和概率模型。其代表分别为布尔模型、向量空间模型和概率模型等。

布尔模型是基于集合理论和布尔代数的一种简单的检索模型。由于集合的概念直观，布尔代数的结构简单，因此布尔模型可以很简单地表示文档和查询，并判断文档与查询间的关系。布尔模型为信息检索提供了一个简单、可操作的框架，这使其可以应用于各类检索系统中。

向量模型通过为查询和文档中的标引词加权以及对检出文档按相似度降序排列来实现文档与查询的部分匹配。

概率模型是基于概率原则的。即给定一个用户查询 q 和集合中的文档 d_j，概率模型试图估计用户找出所需的、与文档 d_j 相关文档的概率，并认为这个相关概率仅依赖于查询和文档的表示。此外，概率模型还假定在文档集合中存在一个子集，用户倾向于把该子集作为查询 q 的结果集。这个理想结果集用 R 来表示，对于用户来说，它能够使总体的相关概率最大，在集合 R 中的文档与查询相关，不在集合中的文档则不相关。

20 世纪 80 年代起，Waller、Kraft、Bookstein、Buell 和 Kraft 等人用部分匹配和词语加权功能来扩展布尔模型，这种方法将布尔查询表达式与向量模型的特征结合在一起[3]。

用关键词集表示文档和查询会形成仅与文档和查询各自真实的语义内容部分相关的描述，由此使文档与查询的匹配是近似的或含糊的。

纵观布尔模型、向量空间模型、扩展布尔模型和概率模型，它们都是以数学为基础，研究代表文档的特征集合、特征向量以及概率的关系。

3 基于场论的信息检索模型

3.1 场论

现代场论的基本思想是把球体对质点的作用看成是球体

基金项目：安徽省自然科学基金资助项目“高精度检索的智能模型研究”(090412054)

作者简介：杨为民(1968－)，男，博士，主研方向：智能信息处理，信息检索，智能软件；李龙澍，教授、博士生导师

收稿日期：2010-12-27 **E-mail：**jsjcx@ahu.edu.cn

激发的引力场对质点的作用，通过势方程求解出势，再进一步求得引力场、引力和引力势能。这种方法在计算上比较简洁，物理思想比较清晰明确。

万有引力场(以下简称引力场)必然存在引力势能。因为两质点间的引力可由公式：

$$\boldsymbol{F}=-G\frac{Mm}{r^2}\boldsymbol{r} \tag{1}$$

表示，它与库伦定律形式类似：

$$\boldsymbol{F}=-\frac{1}{4\pi\varepsilon_0}\frac{q_1q_2}{r^2}\boldsymbol{r} \tag{2}$$

都是与距离平方呈反比的力(q_1 与 q_2 同号为斥力，异号为引力)，并且可计算出这种力的旋度等于0：

$$\nabla\times\boldsymbol{F}=C\nabla\times\frac{\boldsymbol{r}}{r^2} \tag{3}$$

其中，C 为常数，所以这2种力都是保守力，力场是保守力场，这种场存在势及势能。又因为库仑场的电势满足泊松方程，所以万有引力势 U 满足泊松方程。

3.2 文档相关性

对于文档集中的文档 $\boldsymbol{d}_j$，用 $\boldsymbol{w}_j$ 表示文档 $\boldsymbol{d}_j$ 的信息量，进一步，$\boldsymbol{w}_j$ 可以用二元组($\boldsymbol{k}_i$, $\boldsymbol{d}_j$)表示文档 $\boldsymbol{d}_j$ 有标引词 $\boldsymbol{k}_i$，用 w_{ij} 表示 $\boldsymbol{k}_i$ 的权值。用 $\boldsymbol{w}_q$ 表示查询的信息量，同样，$\boldsymbol{w}_q$ 可以用二元组($\boldsymbol{k}_i$, $\boldsymbol{q}$)表示查询中有查询词 $\boldsymbol{k}_i$，用 w_{iq} 表示 $\boldsymbol{k}_i$ 的权值。查询向量 $\boldsymbol{q}$ 表示为 $\boldsymbol{q}=(w_{1q}, w_{2q},\cdots, w_{tq})$，其中，$t$ 是系统中标引词的数目。文档 $\boldsymbol{d}_j$ 的向量可以表示为 $\boldsymbol{d}_j=(w_{1j}, w_{2j},\cdots, w_{tj})$。由此用 t 维向量表示文档 $\boldsymbol{d}_j$ 和用户查询 $\boldsymbol{q}$。

定义1 假定文档的标引词之间是正交的，则文档 $\boldsymbol{d}_j$ 和用户查询 $\boldsymbol{q}$ 间的相互作用 $\boldsymbol{F}$ 为：

$$\boldsymbol{F}=\frac{\boldsymbol{w}_j\boldsymbol{w}_q}{\sum_{i=1}^{t}(w_{iq}-w_{ij})^2}=\frac{\sum_{i=1}^{t}w_{ij}\sum_{i=1}^{t}w_{iq}}{\sum_{i=1}^{t}(w_{iq}-w_{ij})^2} \tag{4}$$

其中，如果文档的标引词不同，即 $w_{ij}\neq w_{iq}$，则标引词之间无关，可令 $w_{ij}w_{iq}=0$，$w_{ij}-w_{iq}=0(1\leqslant i,j\leqslant t)$。计算时间复杂性为 $O(t^2)$。

将搜索结果按文档集中各文档与查询之间的作用力 $\boldsymbol{F}$ 的大小降序排列，即得到文档与查询的相关度。

定义1中采用的是向量空间的文档表示，这样，各标引词之间被假设为是正交的。但实际上文本的标引词之间肯定存在相互的关联，因此，进一步考虑定义有相关性的标引词的文档的作用。

定义2 假定文档的标引词之间存在一定的相互联系 $\boldsymbol{\mu}$，用标引词相关度矩阵 $\boldsymbol{\mu}$ 表示标引词的关系，则文档 $\boldsymbol{d}_j$ 和用户查询 $\boldsymbol{q}$ 间的相互作用 $\boldsymbol{F}$ 为：

$$\boldsymbol{F}=\frac{\boldsymbol{\mu}\boldsymbol{w}_j\boldsymbol{w}_q}{\sum_{i=1}^{t}(w_{iq}-w_{ij})^2}=\frac{\boldsymbol{\mu}_{hi}\sum_{h=1}^{t}w_{hj}\sum_{i=1}^{t}w_{iq}}{\sum_{i=1}^{t}(w_{iq}-w_{ij})^2} \tag{5}$$

其中，$\boldsymbol{\mu}_{hi}$ 为标引词 k_h 和 k_i 之间的相关度。计算时间复杂性为 $O(t^2)$。

由定义1或定义2可知，文档 $\boldsymbol{d}_j$ 和用户查询 $\boldsymbol{q}$ 间的相互作用力也是与 r^2 呈反比的力，并且可计算出这种力的旋度等于0：

$$\nabla\times\boldsymbol{F}=\boldsymbol{\mu}\nabla\times\frac{\boldsymbol{r}}{r^3} \tag{6}$$

其中，$\boldsymbol{\mu}$ 是常数，所以，这种力都是保守力，力场是保守力场，这种场必然存在势及势能。

3.3 文档的相关性分析

根据定义1和定义2，本文用文档所含标引词及标引词的权值表示文档的信息量，文档的相关性使用文档间的相互作用力大小来表示。它与经典的信息检索模型有如下明显的不同：

(1)布尔模型中的相关性是通过判断查询词是否是某一文档的标引词来度量的。如果查询词是标引词，则文档相关，如果不是，则查询与文档不相关，且相关的文档无法进行有效的相关性程度的度量，因此，其结果无法有效地排序。而在定义1和定义2中，如果查询词是文档标引，则文档信息量中包含查询词的信息。进一步可计算出作用的大小用于排序输出。

(2)在向量空间模型中，文档的相关性可以用距离或角度进行度量。距离度量表现在文档 $\boldsymbol{d}_j$ 和用户查询 $\boldsymbol{q}$ 在 t 维空间中对应的点的距离。由于向量空间模型中每个特征词的地位相同，因此可以认为使用距离度量文档相关性实际上是在 t 维空间中以用户查询 $\boldsymbol{q}$ 在 t 维空间中对应的直线为对称轴作一个柱面。用圆柱半径的大小表示相关性的大小。这样对于同一柱面上的文档来说，它们与查询的相关性相同。因此，认为距离度量的向量空间模型的信息检索实质上是对文档的层次聚类。这种层次聚类只能对文档与查询的相关性做出一定程度的区分，无法在同一球面的等相关性文档中进一步分析文档内容。这种距离度量的方法在小语料度的检索中表现不太明显，但在大语料中必然导致检索结果的大量输出。

使用角度度量文档相关性的方法是对归一化后的文档向量与查询向量进行叉积运算：

$$sim(\boldsymbol{d}_j,\boldsymbol{q})=\frac{\boldsymbol{d}_j\cdot\boldsymbol{q}}{|\boldsymbol{d}_j|\times|\boldsymbol{q}|}=\frac{\boldsymbol{d}_j}{|\boldsymbol{d}_j|}\cdot\frac{\boldsymbol{q}}{|\boldsymbol{q}|} \tag{7}$$

式(7)的意义是文档向量对查询向量施加作用或查询向量对文档向量施加作用，其结果是产生一个对 t 维空间原点的2个向量的力距。这实质上并不是度量文档向量与查询向量的关系。

定义1和定义2则直接定义了文档 $\boldsymbol{d}_j$ 和用户查询 $\boldsymbol{q}$ 的信息量间相互作用力的大小，且作用的大小与距离的平方呈反比。所形成的模型是以用户查询 $\boldsymbol{q}$ 为球心作一系列的同心球面，球面半径的大小表示文档相关性的大小。这样既有效解决了单一距离的层次聚类引起大量结果输出的问题，又直接描述了文档 $\boldsymbol{d}_j$ 和用户查询 $\boldsymbol{q}$ 间的关系。

(3)与概率模型、扩展布尔模型和模糊集合模型相比，基于场论的信息检索模型既使用了标引词的词频，又对标引词间的关系进行了模糊化处理，优势明显。

3.4 文档的信息场

基础物理学认为，世界上的物质都可用表征其物质多少的质量进行衡量。带电体的电量可用其所带电荷的多少进行衡量。具有质量或电量的物体会在其周围产生一种物质——场。本文将信息检索中的文档视为物体或带电体，给出如下信息场的定义：

定义3 一个文档信息量可以用其所含标引词的多少进行衡量。具有信息量的文档在其周围产生一种物质，该物质为信息场。信息场的特性是：

(1)任何文档都会在其周围产生信息场。

(2)信息场会对处于其中的文档产生作用。

(3)如果在场力的持续作用下文档的位置发生了变化，则信息场对文档做了功，相应地会发生能量的变化。

在 t 维标引词空间的文档空间中，空间的任一点都可对应于某一文档或查询，其势能可用所有文档对该点的作用乘

以相互间的距离表示。进一步地，除去该点文档的信息量即可得该点的势。

定义 4 对于 t 维标引词空间中的文档 $\boldsymbol{d}_j$ 和用户查询 $\boldsymbol{q}$，查询 $\boldsymbol{q}$ 相对于文档 $\boldsymbol{d}_j$ 的势能 E_{jq}(相对于定义 2)为：

$$E_{jq}=\frac{\boldsymbol{\mu w}_j\boldsymbol{w}_q}{\sum_{i=1}^{t}(w_{iq}-w_{ij})^2}\sqrt{\sum_{i=1}^{t}(w_{iq}-w_{ij})^2}=\frac{\boldsymbol{\mu}_{hi}\sum_{h=1}^{t}w_{hj}\sum_{i=1}^{t}w_{iq}}{\sqrt{\sum_{i=1}^{t}(w_{iq}-w_{ij})^2}} \tag{8}$$

势能 E_{jq} 的大小与文档 $\boldsymbol{d}_j$ 和用户查询 $\boldsymbol{q}$ 呈正比，与间距呈反比。

4 实验与结果

实验使用从中文自然语言处理开放平台(http://www.nlp.org.cn)中下载的文本语料库，该语料库由复旦大学计算机信息与技术系国际数据库中心自然语言处理小组提供，含有环境、计算机、交通、教育、经济、军事、体育、医药、艺术、政治 10 个类别共 2 799 篇文章，训练语料和测试语料基本按照 1：1 的比例来划分，其中，学习语料有 1 402 篇文档；测试语料含 1 397 篇文档。语料库如表 1 所示(表中数据的单位为文本数目)。

表 1 训练及测试语料库

类别	训练语料(1 402)	测试语料(1 397)	合计
环境	100	99	199
计算机	96	96	192
交通	107	107	214
教育	110	110	220
经济	163	162	325
军事	122	121	243
体育	225	225	450
医药	102	101	203
艺术	124	124	248
政治	253	252	505

将基于场论的信息检索与 SMART 系统 11.0 版[4]进行比较。SMART 系统是一个基于向量空间的信息检索系统，其标引词的权值使用词频、文档频、倒排文档和规范化的文档长度计算：

$$\omega_{ik}=\frac{(\ln(tf_{ik})+1.0)}{\sqrt{\sum_{j=1}^{n}[\ln(tf_{ij}+1.0)]^2}} \tag{9}$$

查询的关键词的权值为：

$$\omega_{qk}=\frac{(\ln(tf_{ik})+1.0)\cdot\ln(\frac{N}{n_k})}{\sqrt{\sum_{j=1}^{n}[\ln(tf_{ij}+1.0)\cdot\ln(\frac{N}{n_j})]^2}} \tag{10}$$

其中，tf_{ik} 为标引词(或关键词)t_k 在查询或文档中的词频；N 为文档集的文档总数；n_k 为文档频。

实验从每一类中任选 20 个标引词进行检索。基于场论的信息检索(定义 1 和定义 2，采用同异相关度，j=0.16)和 SMART 系统的平均查准率比较结果如表 2 和图 1 所示。从中可以看出，定义 1 的模型其查准率与 SMART 系统的查准率不相上下，而定义 2 的模型其查准率明显高于 SMART 系统。

表 2 不同检索系统的平均查准率比较 (%)

类别	SMART	定义 1 的模型	定义 2 的模型
环境	15.9	17.1	21.6
计算机	11.8	12.3	14.3
交通	15.8	17.6	21.5
教育	18.9	17.2	27.1
经济	11.6	12.3	16.2
军事	14.3	13.8	17.5
体育	16.5	15.7	18.2
医疗	17.2	17.1	19.2
艺术	16.2	18.5	21.6
政治	20.5	20.2	23.2

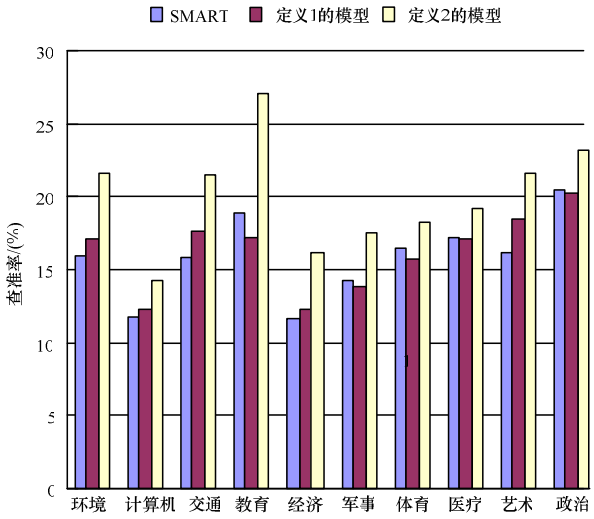

图 1 不同检索系统的查准率比较

5 结束语

本文分析了传统信息检索模型的优劣，提出了基于场论的信息检索模型。该模型借鉴了普通物理学中物体间的引力场作用力和带电体的电场作用力理论，将文档视为具有信息量的信息体。进而以场论的基本特性来描述信息体所产生的信息场及信息场对处于其中的信息体的作用。这样可以从物体之间相互作用的角度阐述文档之间的相关性。实验结果表明，本文模型可以更好地说明文档间的相关性，提高文档检索的查准率。

参考文献

[1] James A. HARD Track Overview in TREC 2005: High Accuracy Retrieval from Documents[EB/OL]. [2010-05-27]. http://trec.nist.gov/pubs/trec14/papers/HARD.OVERVIEW.pdf.

[2] Greengrass E. Information Retrieval: A Survey[R]. Department of Defense, Technical Report: TR-R52-008-001, 2001.

[3] 王秀娟, 郑康锋. 基于文档空间向量距离的查询扩展[J]. 计算机工程, 2009, 35(18): 54-56.

[4] Gerard S. The SMART Retrieval System Experiments in Automatic Document Processing[M]. Englewood Cliffs, USA: Prentice-Hall, 1971.

编辑 张 帆