Computer Science and Application 计算机科学与应用, 2013, 3, 354-359
http://dx.doi.org/10.12677/csa.2013.38062 Published Online November 2013 (http://www.hanspub.org/journal/csa.html)

Algorithm Optimization about Textual Case Retrieval Based on Topic Words*

Zhen Sun[1,2], Hui Yuan[2], Tai Sun[2], Zheng Gong[2], Jie Zhao[2], Lei Tang[3]

[1]Peking University, Beijing
[2]National Administration for Code Allocation to Organizations, Beijing
[3]Chinese Academy of Surveying and Mapping, Beijing
Email: sunzhensh@sohu.com

Received: Sep. 30th, 2013; revised: Oct. 22nd, 2013; accepted: Oct. 30th, 2013

Abstract: Two shortages of Boolean retrieval, ignoring the semantic relations between words and unable to rank the retrieval results in order of importance, were found by analyzing the essence of traditional text retrieval, and in view of which, an improvement of algorithm optimization based on topic words was proposed. Through enriching topic words to structure keywords library, the semantic distance and similarity of keywords were calculated on the basis of semantic retrieval framework. The improved algorithm was applied in the military case retrieval system at last, and then retrieval results were analyzed to detect performance. It is observed that the improved algorithm has a better improvement in both precision rate and recall rate of retrieval.

Keywords: Boolean Retrieval; Topic Words; Semantic Distance; Improved Algorithm; Precision Rate; Recall Rate

基于主题词的文本案例检索算法研究*

孙 镇[1,2], 袁 辉[2], 孙 泰[2], 宫 政[2], 赵 捷[2], 汤 磊[3]

[1]北京大学，北京
[2]全国组织机构代码管理中心，北京
[3]中国测绘科学研究院，北京
Email: sunzhensh@sohu.com

收稿日期：2013 年 9 月 30 日；修回日期：2013 年 10 月 22 日；录用日期：2013 年 10 月 30 日

摘 要：分析传统文本检索方法布尔检索的本质，发现该检索方法存在两个缺点：检索算法忽略了词语之间的语义关系以及不能对检索结果进行重要性排序，针对于此提出利用基于主题词的改进检索算法。通过丰富主题词构建关键词库，在语义信息检索框架的基础上，计算关键词的语义距离和相似度。最后将改进后的算法应用到灾情案例检索系统中，并对检索结果做性能分析，实验证明该算法在文本检索的查准率和查全率上都有较好的改善。

关键词：布尔检索；主题词；语义距离；改进检索算法；查准率；查全率

1. 引言

案例推理(CBR)是最近三十多年来日益发展的区别于规则推理的一种新的推理模式。它是一种重要的基于所积累的知识进行现有问题求解和学习的方式，强调人类对于过去积累的知识经验以及前人的智慧结晶的重视[1]。案例推理的关键在于检索与当前新发

*资助课题：社会管理(微博客)实名备案技术及系统研究(No: 2013 10027)资助课题，国家高技术研究发展计划(No: G1213)资助课题。

案例最相近的历史案例，以直接利用或稍加修改其解决方案来应对当前问题，避免了对类似问题做重复的分析工作，从而大量节省处理问题的时间，因此该推理方法广泛受到国内外研究学者的关注。而如何提高新旧案例匹配相似率，一直是学者们研究的重点。分析以往文献可以发现，过去他们多数是通过比较案例间的属性数值来获取最相似案例，而对于文本的相似计算却研究甚少，另外传统的文本匹配方法已经越来越难满足当前日益增长的信息检索需求。本文基于这一点，提出了基于主题词的文本匹配算法。

2. 传统文本匹配方法——布尔检索

布尔检索是一种简单而常用的严格匹配模型，它定义了一个词组集合来标识文档，该词组被称为标识词组。标识词组对应于文档中的特征项，一般是由训练文档集中的词条或短语组成[2]，如文档的关键词、自由词、作者、片名等。根据调查统计，大部分网络用户在做文档 成检索词组。如此，布尔检索的本质就是将文本匹配转化成词组间的相互匹配。同时运用布尔逻辑运算符“&”(and)和“|”(or)将检索词连接起来形成检索式，再与文档标识词组做逐一匹配，根据是否满足逻辑关系将文档分为两个集合：匹配集和非匹配集。

布尔检索的关键在于如何对文档集中每个文档进行标识，标识的准确性和全面性直接影响到检索系统的查准率和查全率。该方法使得用户在检索时，只需将检索词与标引词进行比较即可，而不必同文档全部内容作逐一比较，具有简单、易理解、易在计算机上实现且检索速度快等优点，故而在很多检索系统中得到应用[3]。

但不容忽视的是，布尔检索同时还存在某些明显的缺陷：

1) 检索词与标识词的比较过于严格，必须要求二者完全一致才算匹配成功。其忽视了词与词之间的内在关系，包括同义关系和包含关系等，如“中国”、“中华人名共和国”和“China”互为同义关系，三者表示同一概念；又如“武器”和“枪械”为概念包含和被包含关系，当用户在搜索“武器”的时候，应当同时能够得到与“枪械”相关的结果集。

2) 检索结果不能根据相关度大小进行重要性排序。布尔检索得到的匹配集中，排在第一位的文档不一定是文本集中最适合用户需求的文档。用户只能按照检索结果的顺序逐一浏览辨别，人工去寻找那些符合自己需求的目标文档，从而降低检索的效率以及用户使用的兴趣。

3. 基于主题词的改进检索技术

3.1. 主题词库的构建

主题词表是文献与情报检索中用以标引主题的一种检索工具。它是由一些规范化的、有组织的、体现主题内容的、已定义的名词术语组成的集合体。但主题词表中固定的词组数目有限，且涉及范围不够广泛，故对其进行扩展，形成更加完善，详细的标引体系，并且存入关系型数据库，形成主题词库，库中表结构如表 1 所示。

主题词库，又称关键词库，是由一系列关键词组成的词库。关键词之间存在着概念包含和被包含关系，可按照树状结构进行呈现，构成关键词树。图 1 就是关键词树的部分示例图。图中每个节点还代表着其同义词，如“枪支”还代表着“枪械”、“枪”、“gun”等。

表 1 中，pid 字段存储树状结构中某节点的父节

Table 1. Table structure of top words library
表 1. 主题词库中表结构

字段名	数据类型	字段说明
id	int(4)	标识
pid	int(4)	父节点标识
nodeName	vachar(200)	节点名
comment	vachar(200)	备注

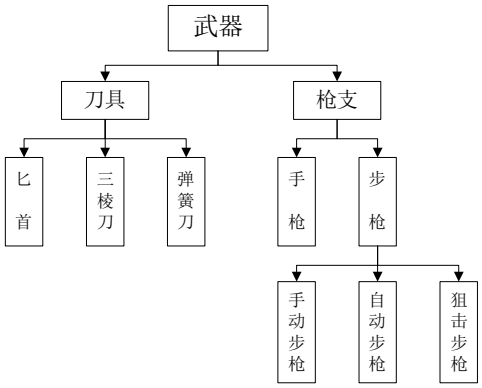

Figure 1. Diagram of keywords tree
图 1. 关键词树示例图

点标识号，对于最顶层节点，pid 取值为“−1”；node Name 字段存储关键词名以及其同义词名，各词之间用符号“\”隔开，且互相为平级关系，先后顺序无影响。

3.2. 基于关键词树的语义距离及相似度

为体现词与词之间的内在联系，引入语义距离概念，通常用 d 来表示。在关键词树中，语义距离指的是关键词连接边的长度。语义距离越小，表示词语概念越接近，反之越远。当语义距离为零时，表明词语一致或互为同义词；当语义距离超过某一临界值时，表明两词没有关联或关联很小[4]。

语义相似度是词语之间内在联系的另一种表示方式，通常用 sim 表示，其与语义距离在数值上成反比关系，最大值为 1，最小值为 0。文本检索的原理就在于寻找文本集中与当前文档语义相似度较大的文档。

3.2.1. 词与词的语义距离和相似度

关键词树中，两个词的连接边长度为他们与其最临近公共父节点距离之和。而子节点与父节点距离实际上指的是他们深度之差，因此可用下列公式表示：

$$d_{A-B}=A\otimes B=\begin{cases}0, & A与B相同或同义\\ D_A-D_B, & A是B的父节点\\ 2D_F-D_A-D_B, & AB最临近父节点F\\ D_B-D_A, & B是A的父节点\\ \infty, & A和B无父节点\end{cases}\quad(1)$$

其中，A、B 是树状结构中的两个节点关键词，F 表示 A、B 的最临近公共父节点，d_{A-B} 指的是 A、B 的语义距离，$\otimes$ 定义为语义距离运算符，D_A、D_B、D_F 分别指 A、B、F 在关键树结构中的深度。根据语义距离公式，可求得相似度：

$$sim_{A-B}=1/d_{A-B}=\begin{cases}1, & A与B相同或同义\\ 1/(D_A-D_B), & A是B的父节点\\ 1/(2D_F-D_A-D_B), & AB最临近父节点F\\ 1/(D_B-D_A), & B是A的父节点\\ 0, & A和B无父节点\end{cases}\quad(2)$$

其中，sim_{A-B} 指的是 A、B 的相似度，注意的是当 A、B 相同或互为同义词时，sim_{A-B} 值为 1。

3.2.2. 词组之间的语义距离和相似度

两个词组的语义距离指的是以其中一个词组为模板，在另一组词里为该模板中的每个词找到距离最近的词，形成临近词组对，再综合求这些词组对的距离。如词组 $P=\{p_1,p_2,\cdots p_m\}$ 和词组 $Q=\{q_1,q_2,\cdots q_n\}$，设其语义距离为 d_{P-Q}。利用向量空间模型法(VSM: Vector Space Model)来计算词组语义距离[5]，分别将 P 和 Q 看成多维空间的两个向量，以 P 为模板，在 Q 中寻找与 P 中各词语义距离最小的词。

$$\begin{aligned}d_{P-Q}&=\min P\otimes Q=\min\begin{bmatrix}p_1\\ p_2\\ \vdots\\ p_m\end{bmatrix}\otimes\begin{bmatrix}q_1 & q_2 & \cdots & q_n\end{bmatrix}\\ &=\min\begin{bmatrix}p_1\otimes q_1 & p_1\otimes q_2 & \cdots & p_1\otimes q_n\\ p_2\otimes q_1 & p_2\otimes q_2 & \cdots & p_2\otimes q_n\\ \vdots & \vdots & \ddots & \vdots\\ p_m\otimes q_1 & p_m\otimes q_2 & \cdots & p_m\otimes q_n\end{bmatrix}\\ &=\begin{bmatrix}p_1\otimes q_x\\ p_2\otimes q_x\\ \vdots\\ p_m\otimes q_x\end{bmatrix},\quad (x=1,2\cdots n)\end{aligned}\quad(3)$$

其中，$P\otimes Q$ 表示向量 P 和 Q 做距离运算，$\min[\ ]$ 表示矩阵每行的最小值，q_x 表示矩阵 Q 中的某个值。

词组间的相似度为各临近词组对相似度的平均值，即：

$$\begin{aligned}sim_{P-Q}&=\sum simd_{P-Q}/m=\sum sim\begin{bmatrix}\min p_1\otimes q_x\\ \min p_2\otimes q_x\\ \vdots\\ \min p_m\otimes q_x\end{bmatrix}\Big/m\\ &=\left(\frac{1}{\min p_1\otimes q_x}+\frac{1}{\min p_2\otimes q_x}+\cdots+\frac{1}{\min p_m\otimes q_x}\right)\Big/m\end{aligned}\quad(4)$$

其中，sim_{P-Q} 代表词组 P 和 Q 的语义相似度，m 指 P 中词的个数，$\sum[\ ]$ 表示矩阵的列求和运算。

3.3. 基于关键词树的语义信息检索框架

图 2 为基于关键词树的语义信息检索模型的基本

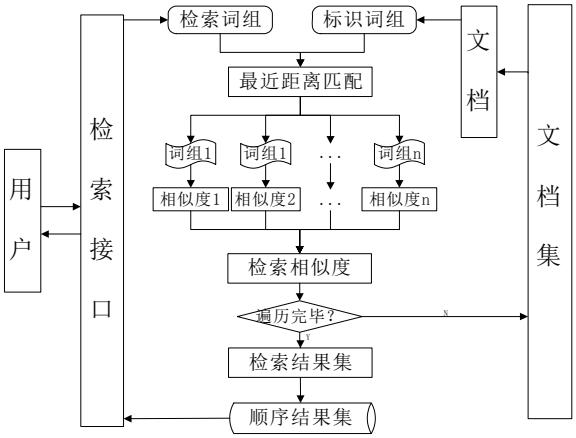

Figure 2. Retrieval framework of semantic information
图 2. 语义信息检索框架

框架。

将用户输入的检索词看成一个数组，用 $P=\{p_1,p_2,\cdots p_m\}$ 表示，m 表示检索词个数；另外将文档集中每个文档的标识词组看成 $Q=\{q_1,q_2,\cdots qn\}$，n 表示标识词的个数。根据上述算法进行文档匹配，步骤如下：

1) 最近距离匹配

为检索词组中每个词寻找距离最近的词，即求 d_{P-Q}。在实际检索中，还涉及到检索词组内部词语的逻辑关系“&”(and)和“|”(or)，所以当存在着某个检索词找不到距离最近的标识词时，对两种逻辑关系做分别处理：

a) 对于形如 $P=p_1\&p_2\&\cdots\&p_m$ 的检索式，则认为该文档不符合检索条件，为非匹配集，直接搜索下一个文档；

b) 对于形如 $P=p_1|p_2|\cdots|p_m$ 的检索式，舍弃该检索词，继续匹配词组中其它词语，但此时 m 的值相应减少 1。既而在相似度计算时，基数也得减少 1。

2) 相似度计算

在对检索词组完成最近距离匹配之后，计算检索词与最近标识词的相似度。再综合各相似度，求其平均值，即为检索词组的最终检索相似度 sim_{P-Q}。设置相似度阈值，将符合要求的检索结果存入检索结果集中，继续匹配下一个文档，直至文档集全部匹配完毕。

3) 排序结果集

结果集中存储着文档检索的相似度，代表着各文档对检索词的符合程度，根据 sim_{P-Q} 值的大小对结果集进行排序，得到顺序结果集。顺序结果集中排在最前面的为最符合解锁条件的文档，依次往下。

4) 输出检索结果

将顺序结果集输出返回给用户，完成检索步骤。

4. 案例检索中的应用

案例检索系统是上述改进算法的扩展与应用。案例检索同普通的检索最大的不同在于，在搜索时输入的不只是检索词，而是一个案例的部分。根据已知的案例内容，去案例库寻找与此最相近的历史案例，从而快速得到解决方案或以此作为参考。

本文以灾情案例为例，一个完整的灾情案例包括“灾情名称”、“灾情种类”、“发生时间”、“发生地点”、“救灾组织机构”、“灾情级别”、“灾情描述”、“灾情影响”、“救灾情况”、“经验教训”等方面。由于是多因子检索，所以此系统在原有算法的相似性检索步骤的基础上，一方面增加了“词库遍历”步骤，其利用关键词树对输入文本进行全文匹配，找出文本中全部关键词以用于检索，解决了用户主动提取关键词的难题以及提高了系统的查全率；另一方面实现了多因子分权重综合匹配，从而大大提高了系统的查准率。

设 SIM 为文档查询的综合相似度，sim_i 表示每个因子的相似度，λ_i 表示该因子在匹配过程中所占的权重，那么

$$SIM=\sum_{i=1}^{n}sim_i\times\lambda_i=sim_1\times\lambda_1+sim_2\times\lambda_2+\cdots sim_n\times\lambda_n$$

$$\sum_{i=1}^{n}\lambda_n=\lambda_1+\lambda_2+\cdots+\lambda_n=1 \tag{5}$$

根据救灾案例的实际情况，将原有的检索框架进行扩展，形成以灾情信息为核心的检索系统框架图，如图 3 所示。

灾情案例检索系统根据上述框架图对用户输入的检索文本进行语义分析，提取灾情种类、地名、救灾组织机构等相关信息，并对其进行语义拓展以得到综合查询结果。如查询“四川地震”相关信息，可在查询文本框中输入“四川省发生地震，红十字会启动紧急救援”，得到如图 4 的查询结果，结果中相似度字段表示与当前查询条件的相关程度。

点击其中一条记录，可查看记录内容。该案例检

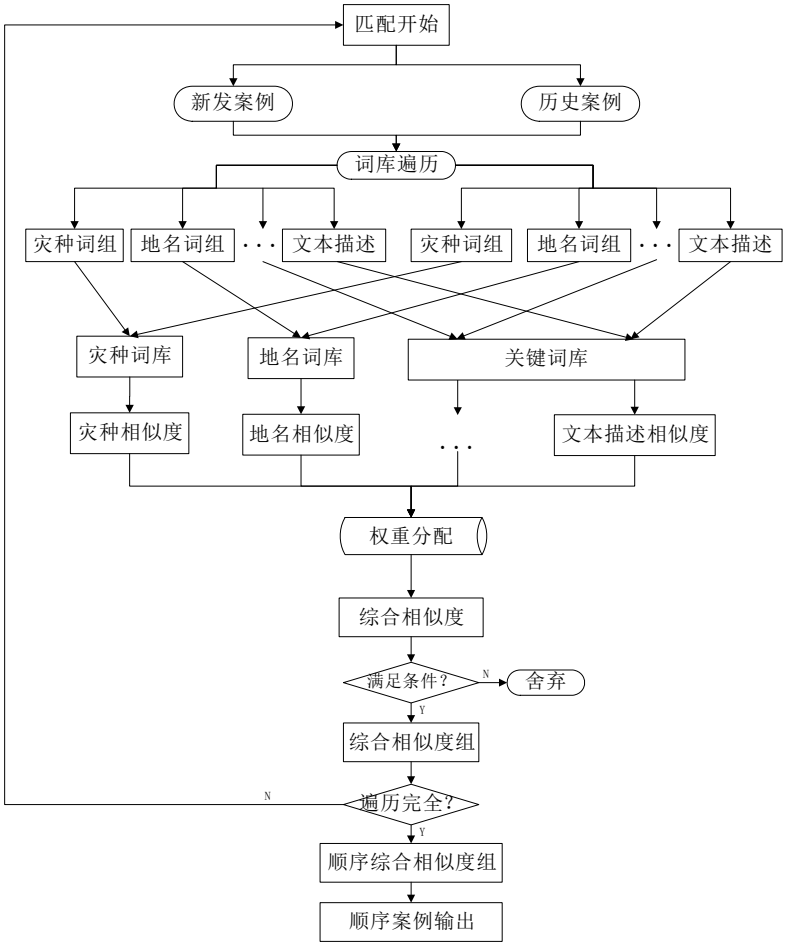

Figure 3. Overall frame
图 3. 总体框架图

序号	灾情名称	灾情种类	发生时间	发生地点	相似度
1	四川雅安7.0级地震	地震	2013年4月20日8时2分	四川省雅安市	100%
2	5.12汶川地震	地震	2008年5月12日14时28分04秒	四川省汶川市	100%
3	四川康定泥石流灾害	泥石流	2009年7月23日	四川省康定县	75%
4	4.14玉树地震	地震	2010年4月14日晨	青海省玉树县	60%
5	1996年云南丽江地震	地震	1996年2月3日19时14分	云南省丽江市	60%
6	1976年河北唐山大地震	地震	1976年7月28日3时42分54.2秒	河北省唐山市	60%
7	广西桂林市全州县滑坡事件	山体滑坡	2011年冬	广西桂林市全州县	37.5%

Figure 4. Retrieved result
图 4. 检索结果

索主要为获取历史案例发生经过、结果和影响等信息。根据此类信息，可预测新发案例的未来发展方向。若结果是有益的，便促进其发展，扩大案例影响范围；反之，及时制定预防措施，甚至是终止其继续发展。

5. 查询性能分析

采用小样本测试，以 1000 个文档作为测试集，选择需要查询匹配的文档进行检索测试。采用查准率 Pre 和查全率 Rec 以及 F-测试值做为评价指标对布尔检索方法和本文所述方法进行评价[6]，计算公式如下：

$$Pre = \frac{\text{检索出的相关文档数}}{\text{检索出的文档总数}} \tag{6}$$

$$Rec = \frac{\text{检索出的相关文档数}}{\text{文档集中文档总数}} \tag{7}$$

$$\text{F-测试值} = \frac{2 \times \text{Pr}e \times \text{Re}c}{\text{Re}c + \text{Pr}e} \quad (8)$$

F-测试值权衡了查全率和查准率，能够较好地反映算法的性能，F-测试值随查全率和查准率的增大而增大，值越大表示查询性能越好，反之越差，最大值为 1，最小值为 0。

考虑到用户平常浏览的只是排在检索前面的一些文档，所以本文在计算查询性能时选择检索结果的前 200 篇文档进行统计，得到查全率、查准率和 F-测试值如下表 2 所示。

从表中数据可以看出，本文对传统的布尔检索算法进一步改进，大大提高了系统检索的查全率以及查准率，其原因就在于本文所述方法考虑了词语语义相关性对检索结果的影响，而不仅局限于文本表面的字符匹配。

6. 结语

本文基于主题词树改善传统检索算法，通过实验分析，在查询结果的完整性和准确性上都有很大的提高，并且在实际系统应用中也取得较好的查询效果。但在关键词树的构建上还存在一些不足，一方面本文全部算法是建立在关键词树的基础上，关键词树构建的越完善，检索效果越好；而另一方面关键词树越复杂，检索的效率会降低，意味着检索时间会增加。所以如何在保证检索效果的基础上，提高检索效率是本文以后努力改进的方向。

Table 2. Result of query performance analysis
表 2. 查询性能分析结果

	查全率	查准率	F-测试值
布尔检索方法	37.5%	64.3%	47.4%
改进检索方法	82.0%	70.8%	76.0%

参考文献 (References)

[1] 严悦, 哈进兵 (2012) 利用 ART 神经网络优化相似案例匹配方法. *信息系统工程*, **3**, 70-74.

[2] 葛继科, 邱玉辉 (2009) 一种基于本体概念语义距离的服务相似度量方法. *计算机科学*, **6**, 181-184.

[3] 王旭阳, 萧波 (2013) 基于概念关联度的只能检索研究. *计算机工程与设计*, **4**, 1415-1419.

[4] 杨健, 赵秦怡 (2008) 基于案例的推理技术研究进展与应用. *计算机工程与设计*, **3**, 710-713.

[5] 杨小平, 丁浩, 黄都培 (2003) 基于向量空间模型的中文信息检索技术研究. *计算机工程与应用*, **15**, 109-111.

[6] Budanitsky, A. and Hirst, G. (2006) Evaluating wordnet-based measures of semantic distance. *Computational Linguistics*, **32**, 13-47.