

多相关主题的主题文本资源相关度算法研究 *

裘江南 王带弟 仲秋雁

摘 要 主题图作为一种新型的信息资源整合工具,在数字资源整合领域有着广泛的应用前景。主题图中资源与主题间的相关度在一定程度上反映了资源与用户查询需求间的相关程度。基于此,本文提出了一种主题文本资源间相关度算法,主要从资源本身及用户动态使用行为两方面深入分析。相关算例试验结果表明,该算法相对稳定,整体上对文本资源排序的效果较好。图 5。表 2。参考文献 11。

关键词 主题图 主题文本资源 相关度算法 数字资源整合

分类号 G202

ABSTRACT As a new tool of information resource integration, topic map is very promising in the field of integrating digital resources. To some extent, the relevance between topics and resources reflects the relatedness between resources and users' inquiries. This paper proposes a kind of relatedness algorithm, which analyzes the resources from the own contents and users' dynastic behaviors. By some related tests, it indicates that this algorithm is relatively stable and can make the resources in good order. 5 figs. 2 tabs. 11 refs.

KEY WORDS Topic map. Resources. Relatedness algorithm. Digital resource integration.

CLASS NUMBER G202

1 引言

主题图作为一种新型的知识组织技术,是信息管理和知识管理之间的桥梁^[1-2],被称为“信息领域的全球定位系统(GPS)”,也是数字资源整合领域的有效工具。它综合了传统索引、文献组织与人工智能等领域知识的优势,可以更加有效地组织信息知识,有利于解决大量无序信息所带来的问题;它提供一个网络信息交换的模型,是沟通信息和知识的桥梁,帮助用户浏览信息资源,并将信息资源层与主题层连接起来,为用户提供基于主题间语义的查询导航。

目前,主题间相关度算法^[3-4]研究较多,而主题与资源相关度研究几乎没有。基于主题图中的资源层绝大多数是文本资源,故文本资源与用户提问主题间的相关度模型的好坏反映了基于主题图的知识导航系统查询结果的好坏。另

外,可以应用于主题与文本资源间对检索结果资源进行排序的传统方法,主要有基于点击率方法^[5]和基于 VSM (向量空间模型, Vector Space Model)的提问词词频方法^[6],这些方法单一,存在很多局限性。为了更好地按用户需求对主题图导航系统检索结果资源进行排序,本文结合该两种方法,提出一种用户提问主题与当前操作主题下的资源间的综合相关度模型。

2 主题图资源分布结构及传统资源排序算法分析

主题图中主题与资源间分布有如下特点:不是所有主题下都拥有资源;同一个主题下有多种类型的资源;同一资源可能从属于多个主题。这些资源可以是异构数据库资源,可以是各种网络

* 本文系国家社会科学基金项目“主题图的语义相关度评价方法研究”(批准号:07CTQ006)成果之一。

类型的资源,也可以是多种信息资源的集合^[7]。主题图下的资源类型主要有:结构化及非结构化的数据库、多媒体(图像、音频、视频)、文本及其他(软件、程序等)一些非结构化或半结构化资源。其中,文本资源是一些非结构化或半结构化数据库,且格式多样(不同格式的电子文档、网页文本及电子邮件等)。由于绝大多数资源是文本资源,检索结果排序的好坏直接影响了整个导航系统查询检索准确率的高低,故这里主要研究文本资源与主题间的相关度算法。主题与资源的分布及分类结构如图1所示:

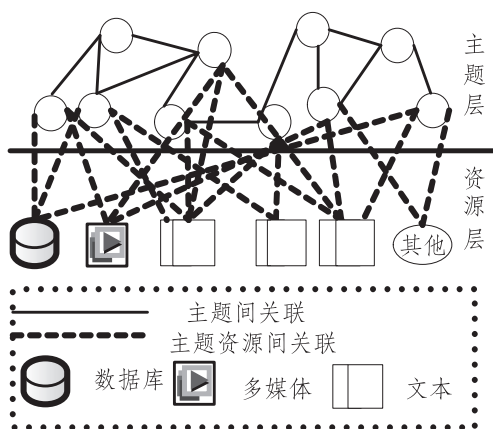


图1 主题图中主题下的资源分布及分类图

主题图技术不仅给予资源精确定位和分类,而且引入了概念间语义关系这个背景知识,反映了概念间的深层语义信息。此外,绝大多数用户是非专业检索人员,其检索提问词往往是模糊的、范围不适当的,不能准确表达其真正的检索需求,因此需要修改检索提问使其逐步与用户真正的检索需求相吻合,而多个主题的提出在一定程度上解决了此问题。传统检索结果资源排序方法主要有基于点击率方法^[5]、基于VSM的提问词词频方法^[6]和网络链接方法^[8],这些方法单一,存在很多局限性。点击率资源排序方法是指查询结果按照点击率从大到小逆序排序,它的最大问题在于,最新的资源由于刚加入数据库,还没有被人利用,因此其排位肯定是在最靠后的,容易被用户忽略,不利于资源利用。后来有学者提出了点击率与时间组合的查询结果组织方法^[3],但结果仍不理想。传统的

VSM是基于提问词集的,它是根据词频建立向量的一种统计模型,需要针对预处理后的文本计算每个词的权重,取权重最高的k个词作为提问词。常用的权重计算方法有布尔权重法、词频权重法、TFIDF权重法^[9]等等,虽对资源排序有一定的效果,但是不同的词汇表示相同的概念,相关的词汇共现表示一个主题等等,均无法表示中文文本中的深层语义信息。近年来出现了一些基于超链接的相关度排序方法,作为基于内容方法的补充,例如PageRank算法及HITS算法^[10]。这两种方法都是基于超文本链接环境的,网页链接动态性强、数量多、数据量大及内容上可能相关也可能无关等特点给网页链接分析带来诸如检索结果覆盖率、网页更新率及排序和网络垃圾等许多问题。本文提出的基于多相关主题的主题文本资源间相关度量化模型分析了影响其相关度量化的主要因素,这些因素分为资源本身特性因素和动态使用因素两大类。多个相关主题TFIDF词频及主题间相关度值的结合,既反映出相关主题间的深层语义关系,又使得文本资源检索结果更全面涵盖用户查询意图。它不仅弥补了传统单一量化方法的不足,而且使资源排序的结果更符合用户查询需求。

3 主题文本资源间相关度算法设计

3.1 相关概念解释

(1) 提问主题:即与用户查询提问词最相关的主题,由于多主题同义现象,该提问主题并不唯一,本算法中只随机选择其中之一。

(2) 相关主题:由在主题图中与提问主题相关联(直接关联或间接关联)的主题及提问主题组成,其数量根据主题图的结构确定。

(3) 操作主题:即为当前需要排序的所有资源从属的某个相关主题,也就是用户目前点击的主题图中最底层的某个主题。

(4) 点击率:本相关度算法中提到的点击率是基于操作主题的点击率。

3.2 主要影响因素分析

根据前人研究的检索结果组织方法(词频、PageRank及HITS、点击率等方法)总结出影响

检索结果组织的因素主要可以分成两类:资源本身特性因素和动态使用因素。其中,鉴于主题图已精确分类及定位其主题下的每个资源,我们把动态使用因素作为重点因素来考虑,语义相关度作为一个附加因素,但是其作用不容忽视,它既反映出文本资源的深层语义信息,而且解决了新挂入资源的排序相关度值的偏低问题。各影响因素详细分析如下:

(1) 文本资源本身特性因素

①相关主题间的语义关系相关度 r 值,即与用户查询提问词相关的主题对于提问主题间关联的相关度值。该相关度值在很大程度上影响了导航的资源与用户需求的匹配程度,且该值越大,相关主题下的资源与用户的检索提问词间的相关程度将越大。②相关主题的 TFIDF 词频,即相关主题在操作主题下的资源中体现的语义权值。值越大,则操作主题下的资源与用户的检索提问词间的相关度值将越大,即该资源与用户需求越相关。

(2) 动态使用因素

文本资源在操作主题下的总点击次数,即用户按照提问词检索资源的过程中对单个文本资源的点击次数(或点击率)。这里文本资源的点击次数是指文本资源在导航系统中在一定的时间内被点击次数,文本资源的点击率是指其点击次数与资源所从属的操作主题被点击次数的比。它更好地反映了文本资源相对其所属操作主题的影响力及其本身的质量。对同一个导航系统而言,资源被点击次数越多,说明该资源与用户查询提问词更相关;对于不同导航系统而言,资源的点击率越大,说明该资源更能满足用户需要、反映的信息更有价值。计算相关度时,为了统一标准,考虑资源的点击率,它是一个 0 到 1 之间的数值。

3.3 算法设计

算法的整体设计思想是:结合文本分类的 TFIDF 公式及向量空间模型,提出一个修正文本语义排序算法,弥补单一算法的不足,提高导航检索的查准率。依据主题图中主题与资源的分布特征,引入反映深层语义信息的主题间相关

度值,结合用户查询词及相关度控制阈值给出相关主题集合,并结合动态使用因素得到算法,基本思想为:对相关主题集合中某个主题来说,出现次数越多,该资源与操作主题越相关;对相关主题集合来说,与提问主题间相关度值越大的相关主题出现的次数越多,该资源与操作主题也越相关;相关主题集合中的相关主题个数出现的越多,该资源与操作主题也越相关。对于导航系统动态使用因素来说,资源相对于操作主题的点击率越大,该资源与操作主题越相关。详细算法设计如下:

定义 1 记 p_{inj} 为相关主题 i 在操作主题 m 下的资源 j 中体现的语义权值(TFIDF 权值因子),其计算公式为:

$$p_{inj} = \frac{pl_{inj} \left| \log \left(\frac{N}{N_k} + \alpha \right) \right|}{\sqrt{\left(pl_{inj} \log \left(\frac{N}{N_k} + \alpha \right) \right)^2}} \quad (1)$$

其中, $\sqrt{\left(pl_{inj} \log \left(\frac{N}{N_k} + \alpha \right) \right)^2}$ 是归一化因子,

N 指操作主题 m 下的文本资源总数, N_k 是指相关主题 i 在操作主题 m 下的文本资源中出现的文本资源总数(计算时,也可以把 N 个相关主题 m 在操作主题下的资源中出现情况用向量空间的形式表示出来,其中的向量都是布尔向量,相关主题在资源中出现便记为 1,否则为 0)。 α 是一个修正参数, α 取值在 0 与 0.1 之间,在文本聚类研究中该值为 0.1,初始实验时取值 0.1,然后其值根据试验结果好坏调整。

定义 2 记 d_j 、 d_m 分别为文本资源 j 在其被挂入时间内,资源 j 在其所属操作主题 m 被点击后被点击次数,操作主题 m 被点击次数,故得到文本资源 j 的相对操作主题的点击率 dl_{mj} 的计算公式为:

$$dl_{mj} = \frac{d_j}{d_m} \quad (2)$$

注:本算法中其取值为文本资源在其所属操作主题 m 下的点击概率值。

定义 3 记 r_i 为文本资源 j 中相关主题 i 与该文本资源所从属的操作主题 m 间的关联相关度值。这里的相关度值的计算采用文献[3]中的相关度算法。

定义4 记 $rel(m, j)$ 为提问主题 m 与文本资源 j 间的相关度值,其计算公式为:

$$rel(i, j) = dl_{mj} \left(\sum_{i=1}^n p_{imj} r_i \right) \quad m = 1, 2, \dots, n(4)$$

3.4 算法的计算机实现

结合主题图导航系统的导航机制及本算法设计得到算法的计算机实现流程图,如图2所示。

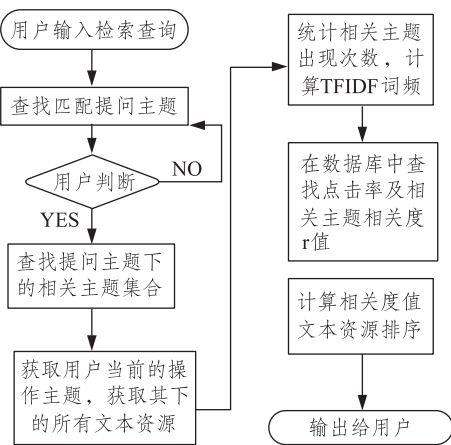


图2 算法实现流程图

4 试验设计与分析

4.1 试验设计

选取台风主题图本体的一个子图,首先提出假设:假设当前用户查询提问为“台风致灾因子”,便得到提问主题“灾害因子”。然后在主题图中查找出其次级主题,分别为次生灾害、风暴潮、大风、暴雨、基础设施、人身安全、城镇建筑、农林牧渔业。接着查找出所有相关主题的集合(相关主题后面括号中的数字为其在主题树中的层次)为{风暴潮(2),大风(2),暴雨(2),洪涝(3),泥石流(3),滑坡(3),人身安全(2),城镇建筑(2),农林牧渔业(2),电力(3),水利(3),交通(3),致灾因子(1)},并运用文献[3]中的主题相关度算法,求得每个相关主题与查询主题致灾因子间的相关度值,这里定义提问主题与它本身的相关度值为1(如图3所示)。

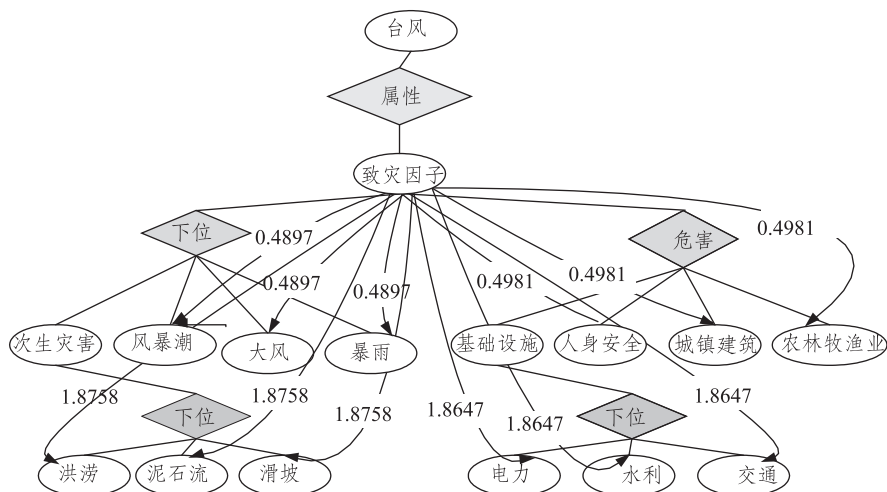


图3 带有提问主题与相关主题相关度值的台风主题树

其次,锁定拟排序的文本资源集合。这里仅锁定其中一个相关主题——暴雨下的19个资源(详见表1),这些资源主要来源于 google 学术搜索及 CNKI,且格式都是 pdf 或 caj 格式。

最后,统计各个文本资源中每个相关主题出现的次数,以备用来计算每个相关主题的主题TIDIF词频因子。

表 1 具体资源列表

标号	资源标题	附注(作者、出处、年代)
1	沿海地区台风灾害系统脆弱性过程诊断与评估——以福建省为例	陈香,灾害学,2007
2	福建台风灾害脆弱性时空趋势评价	陈香,莆田学院学报,2007
3	50年来上海市台风灾害分析及预评估	孟菲等,灾害学,2007
4	热带气旋“碧利斯”与“派比安”灾害成因对比分析	李菁,吴毅杰,何冬燕;灾害学;2007
5	近10年福建台风灾害分析	林炳荣,福建地理,1999
6	江西省台风(热带气旋)灾害特征的分析	谭国良等,中国防汛抗旱,2008
7	2005年福建省台风灾害时空格局与危险性评价	陈香,台湾海峡,2008
8	涝涝灾害对城市的影响分析与防治对策研究	刘永贵等,吉林水利,2004
9	福建台风灾害链分析——以2005年“龙王”台风为例	陈香等,自然科学报,2007
10	福建省台风灾害经济损失变化分析	陈香,莆田学院学报,2007
11	福建省2005年“龙王”台风灾害链巨灾分析	陈香等,莆田学院学报,2007
12	福建省台风灾害风险评估与区划	陈香,生态学杂志,2007
13	迁安市洪涝灾害与防洪减灾对策分析	肖雪梅,农业与技术,2007
14	浅析福建省龙海市防台风减灾策略	曾惠盛,引进与咨询,2006
15	钦州地区水旱风暴雨灾害的特点及其减灾对策	欧柏清,广西水利水电,1991
16	不同登陆地点影响福建的台风灾害时空特征分析	陈香等,亚热带资源与环境学报,2006
17	广东省台风灾害特点及减灾对策	叶雯,灾害学,2002
18	福建省台风灾害及减灾对策研究	陈香等,灾害学,1998
19	基于GIS的台风灾害评估系统设计开发	张斌等,灾害学,2008

4.2 试验结果分析

结合试验设计中已知的相关主题与提问主题间相关度值r集合 {0.4897,0.4897,0.4897,0.4981,0.4981,0.4981,1.8758,1.8758,1.8758,1.8647,1.8647,1.8647,1},用三种算法编程计算相关度值,并对文本资源进行排序。这三种算法分别是:算法1是仅考虑所有资源中的暴雨这个相关主题的TFIDF词频作为资源排序相关度值;算法2是考虑所有相关主题的

TIDIF词频的叠加之和作为资源排序相关度值,此时每个相关主题与提问主题间的相关度值为1;算法3即本试验算法,就是在算法2的基础上,加入多个相关主题与提问主题的相关度值(直接相关度值算中或者间接相关度值)来对资源进行内容语义上的修正,在所有文本资源等点击概率的情况下,以此来计算主题资源相关度值,相关度值计算及文本资源排序详细结果见表2。

表 2 文本资源相关度值

排序	算法 1	资源标号	算法 2	资源标号	算法 3	资源标号
1	0.8429	6	0.1867	19	0.1476	9
2	0.2773	15	0.1856	13	0.1389	18
3	0.2218	9	0.1786	9	0.1334	11
4	0.1664	5	0.1771	14	0.1333	1
5	0.1553	11	0.1747	18	0.1277	17
6	0.1442	14	0.1617	11	0.1274	14
7	0.0998	17	0.1593	1	0.1271	19
8	0.0887	8	0.1441	15	0.1201	6
9	0.0887	3	0.1408	17	0.1159	13
10	0.0776	4	0.1365	4	0.1155	4
11	0.0776	18	0.1310	8	0.1141	5
12	0.0776	1	0.1223	6	0.1138	15
13	0.0665	7	0.1137	16	0.1086	16
14	0.0665	10	0.1133	7	0.1010	7
15	0.0555	12	0.0921	5	0.0968	8
16	0.0444	16	0.0900	12	0.0909	10
17	0.0333	13	0.0871	10	0.0863	3
18	0.0222	2	0.0864	2	0.0831	12
19	0.0111	19	0.0589	3	0.0751	2

通过熟悉台风领域的专家,从 19 个文本资源中评定出符合试验假设中的用户查询需求的相关文本资源,标号集合为: {1 4 5 6 7 9 11 12 13 17 18 19}。然后依据表 2 中的数据及文献 [1] 的 R - 查准率计算方法计算三种算法的 R - 查准率,并画出三种算法比较的直方图,如图 4、图 5 所示。

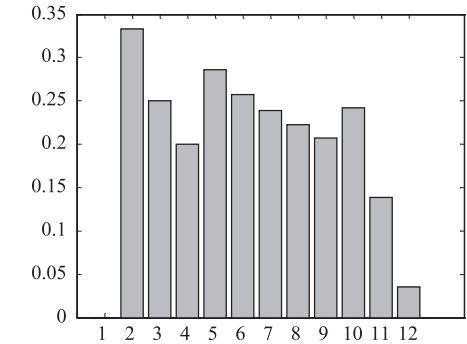


图 4 算法 2 与算法 1 间的 R - 查准率直方图

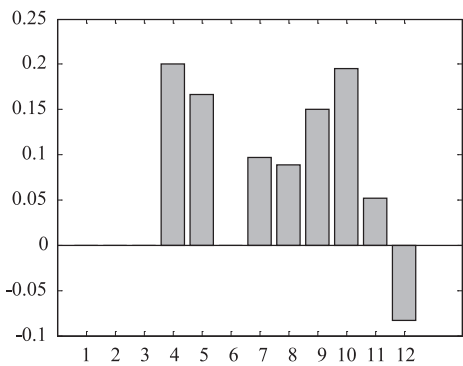


图 5 算法 3 与算法 2 间的 R - 查准率直方图

从图 4 和图 5 可以明显地看出:

第一,考虑多个相关主题的主题文本资源算法 2 与 3 的查准率都要比单考虑相关主题暴雨词频的算法 1 要大很多。仅对 1 个查询文本资源来说两种算法效果相同,而对 11 个查询文本资源来说,算法 2 的 R - 查准率要比算法 1 的

R-查准率高许多,甚至有高达0.34的差距。

第二,既考虑多个相关主题又考虑相关主题与用户提问主题间相关度值的算法3,比仅考虑多个相关主题词频的算法2的R-查准率要略高。如图5中,有4个查询效果相等,7个查询效果较好,1个情况不稳定。这说明由于文本资源所从属的相关主题间的语义关系的加入,文本资源排序结果的确是比不考虑时与用户查询更相关了,但相关主题间相关度算法的好坏影响了主题文本资源间相关度算法的稳定性,即本文提出的主题文本资源间算法对主题间相关度算法的依赖性较大。

第三,从上一点的分析得知:整体来说,本文算法还是比较稳定的,但对相关主题与提问主题间的相关度值的依赖性较大,主题相关度值的好坏直接影响到操作主题与文本资源间的相关度值大小。例如,图5中有一个不稳定情况,算法2的第12个查询相关文本资源的R-查准率高于算法3。这说明内容上与用户查询需求语义上最相关的文本资源不一定使用户满意度高,这一点需要由用户在使用过程中的行为因素来弥补。

综上所述,本文提出的主题文本资源间算法与单一的方法比较,算法相对较稳定,整体上对文本资源排序的效果较好,但仍有改进的空间。

5 结语

随着信息检索技术的飞速发展,基于种种相关性模型的不足,语义计算技术应用于信息检索便成为智能检索的重要方向。基于主题图的相关度算法模型对原有模型做了很大改进,特别是多个相关主题TFIDF词频及主题与提问主题间的相关度 r 值的引入间接地提高了导航系统查全率,挖掘出了文本资源的深层语义信息,使得查询导航结果中反馈给用户的资源与用户在深层语义上最相关。但该模型仍有很多缺陷,没有具体对行为因素进行展开。用户访

问动态行为因素从侧面反映了资源对于用户的价值,可以弥补基于内容相关度算法模型的不足。动态因素虽然较主观,小范围内研究效果并不明显,但仍有很大的研究价值。

参考文献:

- [1] [美]Ricardo Baeza-Yates, Berthier Ribeiro_Neto. 现代信息检索[M]. 王知津,贾福新,郑红军,等,译. 北京:机械工业出版社,2005:52-67.
- [2] 马建霞. 主题图技术在数字化知识组织中的应用研究[J]. 数字图书馆,2004(7):11-16.
- [3] 裘江南,姚永祥,仲秋燕. XTM的主题相关度的量化机制研究[J]. 情报学报,2007,26(3):332-338.
- [4] A Budanitsky, G Hirst. Evaluating WordNet-based Measures of Lexical Semantic Relatedness[J]. Computational Linguistics, 2006,32(1).
- [5] 曾莉红. 基于点击率的全文数据库检索结果组织方法探讨[J]. 情报杂志,2007(6):106-107.
- [6] 徐长青,徐义峰,李舟军. 基于VSM的中文信息检索[J]. 计算机系统应用,2007(4):21-23.
- [7] 马文峰,杜小勇,卢晓惠. 基于知识的资源整合[J]. 情报资料工作,2007(1):51-56.
- [8] 黄晓斌. 网络信息挖掘[M]. 北京:电子工业出版社,2005:63-79.
- [9] 陈庆伟. 文本分类系统的设计和实现[J]. 科技情报开发与经济,2007,17(27):200-201.
- [10] 刘栋,刘希玉,郝婷婷. 基于PageRank和HITS的Web结构挖掘算法研究[J]. 山东科学,2006,19(4):11-14.

裘江南 大连理工大学管理学院副教授。通讯地址:大连市甘井子区凌工路2号。邮编116024。

王带弟 大连理工大学管理学院硕士研究生。通讯地址同上。

仲秋雁 大连理工大学管理学院教授。通讯地址同上。

(收稿日期:2008-12-11)