Wikipedia 中的语义析取

余 旸 林漳希

(北京航空航天大学 经济管理学院，北京 100191) (德克萨斯理工大学 管理学院，德克萨斯 79410)

夏国平

(北京航空航天大学 经济管理学院，北京 100191)

摘 要：维基百科(Wikipedia)现有搜索模块采用关键词匹配方式导致搜索效率相对低下. 为了提高 Wikipedia 中的知识获取效率，提出基于链接分析的词间距算法(TDL，Term Distance based on Linkage). 利用可扩展的计算模型，通过内部链接结构分析发现词簇，并且引入排序和推荐机制. 基于 Wikipedia 2009 年 5 月快照数据的实验表明，TDL 有效增强了 Wikipedia 知识检索的准确性，经由用户评判检验证实 TDL 算法能有效提高用户意图识别度达 7%.

关 键 词：Wikipedia；链接分析；知识发现

中图分类号：TP 182

文献标识码：A **文 章 编 号**：1001-5965(2009)10-1283-04

Extracting thematic communities from Wikipedia

Yu Yang

(School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Lin Zhangxi

(The Rawls College of Business Administration, Texas Tech University, Texas 79410, U. S. A)

Xia Guoping

(School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: The current search module in Wikipedia has low search efficiency due to the search method, which is built on simple keywords matching. To improve the efficiency of knowledge retrieval from the Wikipedia spheres with more accurate links among them, the algorithm named term distance based on linkage (TDL) was proposed. TDL defines a new measure of distance between two keywords, which reorients and organizes those keywords into clusters. It is based on link structure analysis underpinned by computational models. The mechanism of ranking and recommending was imported. The experiment, which based on the snapshot of Wikipedia (May 2009), indicates that TDL would significantly increase the accuracy of knowledge retrieval in Wikipedia and this new algorithm can improve the users' satisfaction by 7% compared with the present one.

Key words: Wikipedia; link analysis; knowledge discovery in databases

目前，包括大部分搜索引擎在内，搜索的实际性能都受限于传统的全文匹配方法，利用关键词和叙词表比照，获得主题词信息，然后依照相似度进行召回. 此类方法对于词间语义相关性的判定存在一定的障碍，因为其中涉及了自然语言理解的问题. 实际上，语义相关性不仅包括词间相似

收稿日期：2008-11-30

基金项目：国家自然科学基金资助项目(70671007)

作者简介：余 旸(1981—)，男，浙江杭州人，博士生，bjyuyang@gmail.com.

性,而且包括词间根据各种语义关系产生的关联性.关于语义相关性的度量研究当前集中于基于大型文集、辞典和结构化本体3种方案.一般而言,第1个方案以需要获得领域相关的大型文集为前提.然而,对大型文集的处理非常困难,得到的结果也并不理想.利用辞典进行相关性计算的局限性在于由专家编撰词典的成本高昂,而且一般只包含小范围的领域和主题,缺乏充分的可扩展性.利用结构化本体,由于其覆盖范围有限,没有涵盖专业性词汇等原因,因此可推广性略显不足.

维基百科(Wikipedia)知识覆盖范围广,描述全面,内容更新迅速,其中每一个词条的解释中,都包括了一系列的相互引用的内部链接.链接指向的词条再向外扩展,形成了一个以词条为节点,链接为边的庞杂知识网,因此其本身就是一个语义处理的理想资源.

基于相关研究[1-10],本文分析Wikipedia词条之间的超链接结构信息并设计词条相关性算法,在此基础上对所有词条进行相关度排序并引入推荐评价机制,以构造一个自适应的主题搜索模型.在实验评估中,将此方法用于英文版Wikipedia采用人工评价的方法测评此模型下的查询结果用户搜索满意度,并和当前方法产生的结果列表进行比较.

1 问题描述

假设使用全文匹配检索方法,在数量庞大的结果中,很大一部分并不是用户真正关心的信息,同时可能也没有包括一些重要的相关信息.如图1中,假如用户输入查询关键字"中华人民共和国",由自然意群角度判断,"1949"和"北京"应当

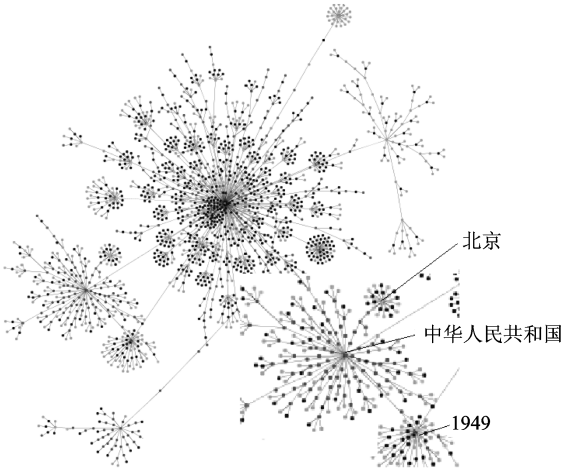


图1 Wikipedia内部链接结构

属于相关词条,然而使用全文匹配检索方式并不能召回这样的相关信息.事实上,在一个类数据库视图中,用户查询的词条被认为等同或属于另一个词条,并且和其它的词条存在着或强或弱的联系.理想的查询应当是以用户意图判定和词群划分的质量为基础的.因此,本文的研究问题可以定义如下:

1) 在Wikipedia的知识空间中辨别语义关系的强弱;

2) 侦测出缺失的强语义链接,从而尽可能完整地析取资源中的语义关系.

2 基于链接结构的词间距算法

本节将介绍基于链接结构的词间距算法(TDL,Term Distance based on Linkage),算法描述伪代码如下.本文将利用此算法识别并排序语义信息.

Input: $G(t)=(V,E)$// 带权有向图

Output: TDL矩阵 $D(N*N)$// 共有 N 个节点

```
For (i=0;i<N+1;i++)// 计算 TDL
{
 For (j=0;j<N+1;j++){
   Calculate D(i,j){
    L_i, L_j: select the level of V_i ∈ {V} (Breadth-First-Search)
    x_ij: formula (2)// 计算系数 x
    y_ij: formula (3)// 计算系数 y
    D'_ij: formula (4)// 计算 TDL 值 D'
    }
  }
}
    For (k=0;k<N+1;k++){ // 分别提取相关词群
        Sort the D'_k}
```

Wikipedia中的独立词条通过建立指向其它词条的链接发生词间联系.因此链接对于动作发出者记为出链(OutLink),对于接受者(链接指向的词条)记为入链(InLink).

为了获取词间的链接建立行为以及词群的形成动态,构造一个时变带权有向图(time-dependent weighted directed graph) $G(t)=(V,E)$,用于描述某一时点的全集状态.其中,V 代表词条集合;E 代表所有指向关系.

将实际的词条链接抽象成为图2所描述的一般模型.可以想象,关系紧密的词条之间通常蕴涵

着更强一些的语义联系[11-13]. 如上文,以“中华人民共和国”和“北京”为例——对应为顶节点,2 个词条共同指向的词条数量(中间层及叶子节点数)和词条之间的交互出入链数量(单向和双向链接数)较大,可以由此推断两者之间存在着较强的语义相关性. 再如“北京”和“巴黎”之间几乎没有出入链关系,因而据此推断出其弱语义相关. 进一步将之模型化,

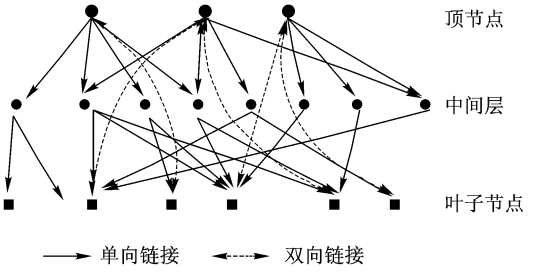


图 2　词条树一般模型

假设 1　词条树中同层链接对于词间距贡献相同.

假设 2　词条树中距离顶节点层数与词间距贡献量成反比.

设定算法计算量如下:

1) 共有链接数. 2 个词条的出入链关系越多,那么认为其语义关系越强烈. 对于 OutLink 和 InLink 同时存在的情况,本文对之分别处理,每个链接分别计数.

2) 链率. 为了均衡统计中的偏斜,在 TDL 计算中,首先将链接数标准化. 使用出入链与词条自节点数的比值作为计算量.

2 个词条链接树 a 和 b,其中 a 有 I 层子节点,每层节点数表示为 A_i,b 有 J 层链接,每层的节点数表示为 B_j,L_{ij} 为词条 a 的第 i 层对词条 b 第 j 层的出链. 那么 a 对于 b 的 TDL 值 D_{ab} 有:

$$D_{ab}=\sum_{i=1,j=1}^{i=I,j=J}\frac{L_{ij}}{A_i+B_j}\frac{}{I+J} \tag{1}$$

考虑到如果 2 棵树在层次以及同层节点 2 个方面都比较接近,那么它们之间的相似度也越高.

因此,引入系数 x,与层次数的差负相关;t 为常数;令

$$x=t\sqrt{1-\frac{|I-J|}{I+J}} \tag{2}$$

同理,引入系数 y,和同层节点数差负相关;k 为常数;令

$$y=k\sum_{i=1,j=1}^{i=I,j=J}\sqrt{1-\frac{|A_i-B_j|}{A_i+B_j}} \tag{3}$$

将 2 个系数引入公式(1)有:

$$D'_{ab}=D_{ab}\times x\times y \tag{4}$$

3　实验结果与分析

本节中,将 TDL 应用于英文版 Wikipedia 2009 年 5 月快照数据并分析实验结果.

3.1　数据集描述

首先,析出文集包含所有的内部链接. 数据集包括独立词条数 2 140 512 个,链接数 122 847 716 个.

随机抽取 500 个词条作为实验样本并对每个词条按照链接关系构造自索引关系数据库. 共包括词条 1 055 781 个,平均出链为 51.7,词条树平均层次为 3.4. 格式如表 1 所示.

表 1　词条树存储结构

初始 ID	初始节点	目标 ID	目标节点
…	…	…	…
131	People's Republic of China	25 512	Beijing
131	People's Republic of China	368 723	Country
…	…	…	…
25 512	Beijing	1 416 219	Ancient Capitals of China
25 512	Beijing	365 411	Shanghai
368 723	Country	13 489	Government

3.2　实验结果

对于抽取以及计算结果按照其 D′值排序取前 4 个词条得到推荐列表(t, k 分别为 0.5),并采用人工评分的方法判断此结果和目前搜索方法产生的列表对于用户满意度的变化情况. 将此结果与目前搜索得到的推荐结果进行人工评分对比. 评分标准为:按照推荐条目符合其搜索需求的程度由高到低分为 3,2,1,如不相关判为 0.

表 2 列举了部分推荐排序与当前搜索结果对比. 表 3 为 TDL 算法与现推荐列表的人工测评分数统计结果,此结果通过置信度为 0.05 的 t 检验. 分析表明,采用 TDL 算法对用户搜索意愿的识别度要高于目前搜索算法,满意度由原来的 1.82提升至 1.95,提高幅度为 7%. TDL 与现搜索算法用户满意度对比箱图见图 3.

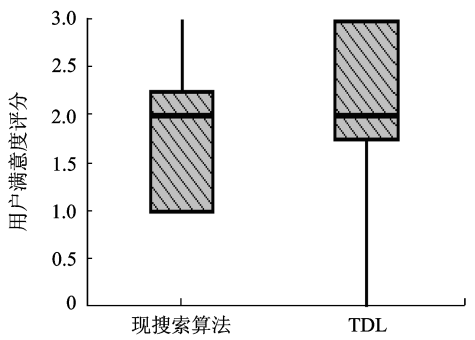


图 3　TDL 与现搜索算法用户满意度对比

表2 TDL与当前搜索结果对比(部分)

搜索词条	排序结果			
	TDL	满意度	当前结果	满意度
Internet	Basic Internet	3	Outline of the Internet	1
	Computing	2	Demographics of the Internet	1
	Information Technology Management	2	Digital rights	2
	Software Engineering	2	Information continuum	1
Agriculture	Basic Agriculture	3	Aeroponics	1
	Organic Gardening and Farming	1	Agricultural economics	2
	Organic Gardening	2	Agricultural marketing	2
	Sustainable Agriculture	1	Agro ecology	1
Health	Basic health	3	Environmental health	0
	Health	2	General fitness training	1
	Medical	2	Health care	3
	Psychological	2	Hygiene	2
Mathematics	Basic Discrete Mathematics	2	Dyscalculia	0
	Basic Mathematics	2	Iatromathematicians	1
	Mathematical Logic	2	Logics	2
	Set Theory	1	Mathematical anxiety	2
Nutrition	Basic Nutrition	0	Balanced Eating	2
	Biochemistry	2	Biology	1
	Nutrition	2	Dangers of poor nutrition	2
	…	…	…	…

表3 TDL与当前搜索结果相关度得分对比

算法	样本数	均值	均方差	标准误差
当前搜索结果	2 000	1.82	0.70	0.016
TDL	2 000	1.95	0.79	0.017

4 结束语

本文分析Wikipedia内部链接信息,设计基于链接结构的词间距算法TDL.通过人工评分机制和当前搜索结果进行用户搜索满意度对比表明,TDL对于用户满意度有明显的提高.

但是,TDL实验结果呈明显左偏和右尾消失,且方差较大.说明在TDL算法产生误判而导致推荐无效(得分偏低甚至为零)的情况增多.因此,在下一步研究中将进一步改进排序算法,改善算法在甄别弱相关语义方面的性能,并计划考虑结合内容角度的词条相似度测量来提高搜索精度.另外,目前采用的人工评分机制主观性偏大且区分度不高,需要进一步完善.

参考文献(References)

[1] Markus K, Denny V, Max V. Wikipedia and the semantic web-the missing links[C]//Walter V. Wikimania 2005. Frankfurt am Main, Germany: Association for Computing Machinery Press(ACM),2005:117-125

[2] Max V, Markus K, Denny V, et al. Semantic Wikipedia [C]//Leslie C. WWW2006. Edinburgh, Scotland: Association for Computing Machinery Press(ACM), 2005: 265-274

[3] Shawn D A. Structure helps a Wiki navigate[C]//Mohammad A. WebDB 2005. Arlington, VA: AAAI Press, 2005: 97-108

[4] Natalia K. Automatic ontology extraction for document classification[D]. Saarbrücken: Computer Science Department, Saarland University, 2006

[5] Daniel K. WikiSense-mining the Wiki[C]//Walter V. Wikimania 2005. Frankfurt am Main, Germany: Association for Computing Machinery Press (ACM),2005:254-276

[6] Chakrabarti S. Data mining for hypertext: A tutorial survey [C]//Usama M F. SIGKDD Explorations. Cambridge, Massachusetts: MIT Press, 2000:113-125

[7] Jakob V. Measuring Wikipedia[C]//Peter I. ISSI 2005. Stockholm, Sweden: Karolinska University Press, 2005:21-36

[8] Francesco B, Roberto B. Network analisis for Wikipedia [C]//Walter V. Wikimania 2005. Frankfurt am Main, Germany: Association for Computing Machinery Press (ACM),2005:334-367

[9] Sergey B, Lawrence P. The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks and ISDN Systems, 1998, 30(1-7):107-117

[10] Jon K. Authoritative sources in a hyperlinked environment [R]. Technical Report RJ 10076, IBM, 1997

[11] Fernanda B, Martin W, Kushal D. Studying cooperation and conflict between authors with history flow visualizations[C]//Brian B. SIGCHI 2004. Vienna: Association for Computing Machinery Press (ACM),2004:575-582

[12] Salton G. Automatic text processing: the transformation, analysis, and retrieval of information by computer[M]. New York: Addison-Wesley, 1989:11-17

[13] Broder, Henzinger M. Information retrieval on the web: Tools and algorithmic issues [M]. Austin: Addison-Wesley, 1998:112-145

(编 辑:文丽芳)