

基于动态匹配的主题相关度判定算法

郑 敏

(武警工程学院 通信工程系, 陕西 西安 710086)

摘 要: 本文在基于元数据的判定方法和基于网页内容的语义判定方法的基础上, 借鉴哈弗曼编码的思想, 提出了一种基于动态匹配的主题相关度算法. 该算法通过分析输入数据的类型来采用相适应的判定方法. 实验结果表明, 从准确率和速度相统一的角度看, 本文提出的主题相关度算法要优于现有算法.

关键词: 主题相关度; 主题型搜索引擎; okapi 公式

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2012)01-0105-04

Evaluation Arithmetic of Topic Relevance Based on Dynamic Matching

ZHENG Min

(Department of Communication, Engineering College of Armed Police Force, Xi'an 710086, China)

Abstract: An evaluation arithmetic of topic relevance based on dynamic matching is presented in this paper, which is based on evaluation of metadata and evaluation of semantics based on web content. The computing model can adjust automatically based on the type of inputted data and is able to settle in great part the contradiction between process speed and accuracy rate. The experimental results demonstrate that the computing model presented in this paper perform better than present approaches.

Key words: topic relevance; a topic-driven search engine; okapi formula

1 引言

随着信息化的高速发展, 信息量正以指数规律迅猛增长, 互联网已经成为人类最重要的海量信源, 如何有效地提取并利用这些信息成为一个巨大的挑战^[1]. 人们可以通过像 Google 和百度这样的通用搜索引擎来检索信息. 然而, 专业领域的用户使用通用搜索引擎来检索信息时, 返回的大多数网页对他们而言是没有用的^[2]. 因此, 面向专业领域的主题型搜索引擎的研究是非常必要的并且已经成为了搜索引擎研究的一个热点问题. 主题型搜索引擎的关键技术之一是判定目标网页的主题相关度, 即判定目标网页与指定主题的相关程度^[3].

2 网页主题相关度的判定方法

网页主题相关度判定是用来判定某一主题与带判定网页的相关程度, 一般用相关系数来表示. 网页相关度判定是主题型网页信息检索中的基础技术之一, 现有的判定方法主要有以下几种^[4].

(1) 基于元数据的判定方法

网页的设计者会在页面的元数据部分填入该网页的概要信息, 例如该网页的作者、关键词以及概要内容等. 该方法通过分析元数据来判定网页与主题的相关性. 这种方法耗时少, 但有很大的局限性, 即如果网页设计者没有提供充足的元数据, 或者提供的元数据与网页的实际内容不符, 那么这种方法的准确率就无法得到保证.

(2) 基于链接文本的判定方法

大量的统计结果表明,链接的链接文本与链接所指的网页内容具有很强的相关性,故可以通过分析链接文本来预判定目标网页的主题相关度。

(3) 基于网页内容的判定方法

统计结果表明,对网页主题相关性判定最准确的方法还是要针对网页内容进行语义判定,但是和其他两种方法相比,该方法的耗时要大的多。

3 基于动态匹配的主题相关度算法

受哈弗曼编码思想的启迪,本文在网页切分和网页主要内容块发现的基础上^[5],先采用基于元数据的判定方法进行分析,然后结合网页内容的特征,依据分析结果来动态匹配不同形式的基于网页内容的语义判定方法,最后综合计算网页的主题相关度。

3.1 对 okapi 公式的改造

实验表明,在已有的权重计算函数(如布尔权重函数、TF-IDF 权重函数等等)中,Okapi 权重计算公式是所表示的文本向量空间中最为合理的权重函数^[6]。

因为 okapi 公式是用于计算文本中某词项的权重,而本文要计算的是主题中某词项的权重,因此文中对 okapi 公式进行了改进,提出了一种针对主题的权重计算公式 okapi FS(for subject),如公式(1)所示。

$$a_i = \frac{\sum_{j=1}^n tf_{(i,j)}}{0.5 + 1.5 \frac{\frac{1}{n} \sum_{j=1}^n dl_j}{avg - dl_N}} \log \left(\frac{N - df_N + 0.5}{df_N + 0.5} \right) \quad (1)$$

$$avg - dl_N = \frac{\sum_{j=1}^N dl_j}{N}$$

其中, a_i 为某一主题中词汇 i 的权重; $tf_{(i,j)}$ 为某一主题的文本训练集中,词汇 i 在文档 j 中出现的次数; df_N 为训练集中包含有词汇 i 的文档数量; dl_j 为文档 j 的长度; $avg - dl_N$ 为训练集中文本的平均长度; N 为训练集中所包含的文本的数量; n 为某一主题所包含的文本数量。

3.2 非线性主题相关度计算公式

研究发现,就词频而言,主题相关词在同一篇文章中出现次数越多则越显得重要,但是,不能简单的认为主题相关词的重要程度与其在文献中出现的次数成正比,这种简单地用线性函数表示两者之间关

系的方法与实际的情况有很大的差别^[6]。

为了接近实际,引入一个调节系数 c 来控制函数。本文提出了一种非线性函数(图1):

$$f(n) = \begin{cases} n/(n+c), & \text{当 } c \neq 1 \\ n/(n+1), & \text{当 } c = 1 \end{cases} \quad (2)$$

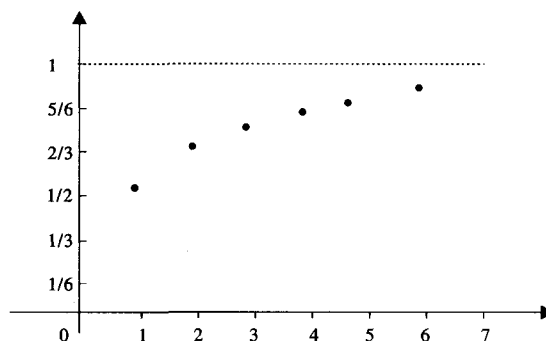


图1 $f(n)=n/(n+1)$

3.3 针对文本的主题相关度算法

基于对上述问题的研究,本文构造了针对普通文本的相关度的算法,如公式(3)所示。

$$Y = \frac{[a_1 f(x_1) + a_2 f(x_2) + \dots + a_n f(x_n)]}{L}$$

$$f(x) = 1 - \frac{c}{x+c} \quad (3)$$

其中, a_i 为第 i 个关键词的权重; x_i 为第 i 个关键词的词频; L 为文本的关键词数; c 为调节系数; Y 为文本的主题相关度。在对目标文本进行判定时,先去除停用词,分析出公式(3)需要的所有参数,这样就可以计算出目标文本的主题相关度了。

3.4 针对网页的主题相关度算法

本文在分析了基于元数据的判定方法和基于网页内容的语义判定方法的基础上,再结合两种方法的基础上进行改进,提出了一种新的基于动态匹配的主题相关度方法,其流程如图2所示。

3.4.1 基于元数据的判定方法

基于元数据的判定方法主要是根据 HTML 的 meta 标签中的标题,关键词,网页描述三个属性所提供的信息进行分析来计算网页的主题相关度的^[7]。如公式(4)所示。

$$Y_1 = W_t * S_t + W_k * S_k + W_d * S_d \quad (4)$$

$$W_t + W_k + W_d = 1,$$

$$\text{且 } W_t \geq 0, W_k \geq 0, W_d \geq 0.$$

其中, Y_1 为用基于元数据的判定方法计算出的网页主题相关度; W_t 、 W_k 、 W_d 分别为标题、关键词、网页描述的权重值; S_t 、 S_k 、 S_d 分别为标题、关键词、网页

描述的主题相关度得分,可以通过公式(4)分别计算得出.

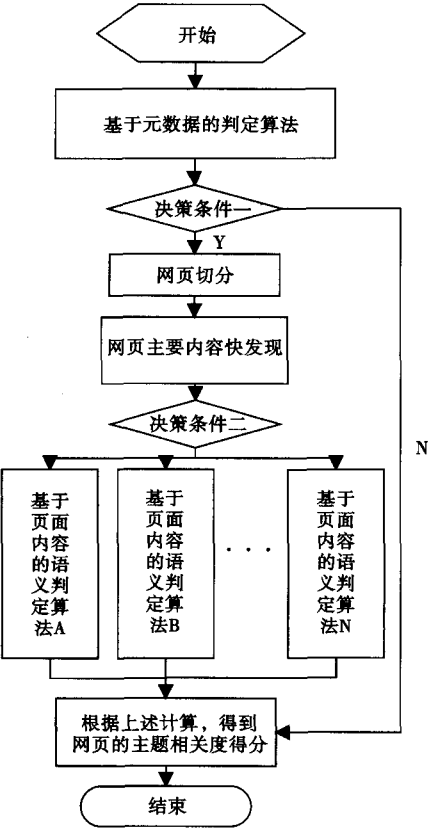


图 2 基于动态匹配的主题相关度算法流程图

3.4.2 关于决策条件一

决策条件一(见图 2)是对网页的主题相关度 Y_1 的判定.使用基于元数据的判定方法对某一网页进行分析,如果 Y_1 值大于阈值 V ,则进一步对网页相关度进行分析;否则,认定此网页与主题不相关,不作进一步分析.

3.4.3 关于决策条件二

决策条件二(见图 2)是在针对网页的主题相关度 Y_1 、篇幅以及网页中文本内容及其相对位置进行分析后,动态地匹配基于内容的判定方法.决策条件二的设计思路是以下面三个理论依据为基础的.

(1) Top n 法则是搜索引擎领域中的公认法则.如果对 Top n 法则进行扩展可以得出推论,即如果网页与主题的相关性越强,则越应该精确地计算其主题相关度;反之,如果网页与指定主题的相关性很小,那就没有必要计算其主题相关度了^[8].

(2) 经研究,从判定主题相关度的角度看,出现在题目中的词项要比出现在正文中的词项权重大;出现在首段末段的词项要比出现在其他段落中的词项权重大;出现在某一段落中的首句末句中的词项

要比出现在段落的其他部分的词项权重大^[8].

(3) 假设两个网页分别为 Web1 和 Web2, Web1 的词项数是 Web2 的 N 倍.如果采用相同的判定算法,则分析 Web1 的复杂度大于等于 Web2 的 N 倍^[9].

设 A 方法为基于元数据的判定方法,B 方法为基于网页内容的判定方法, L 为网页文本的长度, Y 为采用 A 方法计算出的主题相关度,依次假设 B_1 、 B_2 、 B_3 分别代表采用扫描全文、扫描首段末段以及其他段落的首句末句、扫描各个段的首句末句的 B 方法.决策条件(2)的动态匹配方法如表 1 所示.

表 1 决策条件(2)的内容

	$L < L_1$	$L \geq L_1$
$Y \geq Y_2$	B_1	B_2
$Y_1 \leq Y < Y_2$	B_2	B_3
$Y > Y_1$	B_3	不处理

3.4.4 计算 Web 网页的最终主题相关度得分

综上所述,Web 网页的主题相关度的最终计算方法如公式(5)所示.

$$Y = W_m Y_1 + W_c Y_2 \tag{5}$$

$$W_m + W_c = 1, \text{ 且 } W_m > 0, W_c > 0$$

其中, Y 为 Web 网页的主题相关度的最终得分; W_m 为 Y_1 的权重值; W_c 为 Y_2 的权重值; Y_1 为基于元数据的判定方法计算出的网页主题相关度值; Y_2 为基于内容的判定方法计算出的网页主题相关度值.

4 实验结果与分析

针对军事信息的主题型搜索引擎,首先对该系统的使用日志进行统计,统计结果如表 2 所示.

表 2 Web 网页与军事主题相关的可能性

	高相关	低相关	不相关
网页处于各范围可能性	1%	8%	91%

以该统计结果为依据来设计实验数据.分别取相关度较高的、相关度较低的以及不相关的网页 10 篇、80 篇、910 篇.对这些网页分别采用基于元数据的判定方法、基于网页内容的语义判定方法以及本文提出的网页相关度判定方法进行实验.实验结果如表 3 所示.从实验结果可以看出,综合考虑耗时性和准确性,本文提出的算法最优.

表 3 不同方法的准确率与耗时的对比

	耗时(ms)	准确率
基于元数据的判定方法	14857	45.6%
基于网页内容的语义判定方法	1876553	97.3%
本文方法	43449	95.4%

另外,使用文中提出的方法,分别对相关度高的、相关度低的以及不相关的网页进行了实验,实验结果如表 4 所示. 可以分析出:主题相关度越高的网页其相关度判定的准确率也越高,由于用户一般不会关注排名 50 之后的信息,因此,网页主题相关度低影响到判定准确率的问题几乎不会对用户造成影响.

表 4 本文方法针对不同类型数据的准确率

	主题相关度高的 Web 数据	主题相关度低的 Web 数据	不相关的 Web 数据
准确率	98.8%	81.5%	45.7%

5 结束语

高准确率和低耗时是相互矛盾的,本文提出的基于动态匹配的主题相关度算法在很大程度上减少了这种矛盾. 实验结果表明:从准确率和耗时相统一的角度,本文提出的方法达到了较好的效果.

参考文献:

[1] 中国互联网络信息中心. 第二十六次中国互联网络发

展状况统计报告[EB/OL]. [2011-05-28]. Retp://re-search. cnnce. cn/html.

[2] 王继成,萧嵘,孙正兴,等. Web 信息检索研究发展[J]. 计算机研究与发展, 2001,38(2): 187-193.

[3] F Menczer, G Pant, P Srinivasan, et al. evaluating Topic-Driven Web Crawlers [C]// Proceedings of the 24 Annual International ACM/SIGIR conference, New Orleans, USA; ACM, 2007.

[4] Cutler M, Shih Yung-ming, Meng Wei-yi. Using the structure of HTML documents to improve retrieval [C] // Proceedings of the 11st IEEE Conference on Tools with AI, CA, USA; IEEE, 2009;406-409.

[5] 韩先培. 基于布局特征与语言特征的网页主要内容块发现 [J]. 中文信息学报. 2008(1):15-21.

[6] 张天广. 基于 Internet 的银行竞争情报收集系统的研究与实现 [D]. 西安: 西北大学, 2009.

[7] Yi J, Sundaresan N. Metadata based web mining for relevance [C]// Proceedings of the In IEEE 2000 International Database Engineering and Applications Symposium (IDEAS'00) Yokohama; IEEE, 2010: 113-121.

[8] Bing Liu, Chee Wee Chin, Hwee Tou Ng. Mining topic-specific concepts and definitions on the web [C]// Proceedings of the 12th international conference on World Wide Web, NY, USA; ACM, 2003: 251-260.

[9] Sergey Brin. Extracting patterns and relations [J] Computer Science Department Stanford University, 2009: 172-183.

作者简介:

郑敏女, (1982—), 博士研究生, 讲师. 研究方向为搜索引擎复杂系统建模研究.

(上接第 104 页)

参考文献:

[1] 胡亮, 袁巍, 于孟涛, 等. 单向性策略与 AES 密钥生成算法的改进 [J]. 吉林大学学报:工学版, 2009, 39(1): 138-141.

[2] 景为平, 徐晨, 陈海进. AES 密码算法的结构优化与实现 [J]. 微电子学与计算机, 2007, 24(2): 36-38.

[3] 吴文玲, 冯登国. 分组密码工作模式的研究现状 [J]. 计算机学报, 2006, 29(1): 22-25.

[4] 闫茂德, 纪志强, 贺显曜, 等. AES 与 ECC 混合加密算法的无线数据通信系统设计 [J]. 微电子学与计算机, 2007, 24(7): 135-138.

作者简介:

杨小东 男, (1985—), 硕士. 研究方向为加密技术与无线通信.

王毅 男, (1966—), 博士, 博士生导师. 研究方向为电力电子及电气传动.