基于元搜索的信息检索模块的研究和实现①

李 欢，刘 倩，张 英

(石家庄军械工程学院 计算机工程系，石家庄 050003)

摘 要：在对中文问答系统分析的基础上，提出了多信息融合相似度的计算方法，使得信息检索模块中的候选结果集能够按照相似度高低进行排列，为后续的答案提取模块提供了依据。实验表明，该方法提高了页面检索的查准率，更好的符合了用户的实际检索需求。

关键词：问答系统；元搜索；信息检索；多信息融合；相似度

Research and Implementation of Information Retrieval Based on Meta Search Engine

LI Huan, LIU Qian, ZHANG Ying

(Department of Computer Engineering, Ordnance Engineering College, Shijiazhuang 050003, China)

Abstract: The paper introduces a method of multi-information similarity, based on the analysis of the Chinese Question-Answering System. And then the upcoming module of answer extraction could use the result set of candidacy from information retrieval to satisfy user's needs by rank of result set. The experiment shows that this method improves accuracy in retrieval of Web pages.

Keywords: question-answering system; meta search engine; information retrieval; multi-information; similarity

1 引言

传统的搜索引擎一般采用的是集中方式，它们利用一种称为网络机器人[1,2](Robot)的自动化程序来遍历互联网，对能搜索到的文档生成全文索引，供用户检索。这种方式最大的弊端是覆盖度不高，根据专家的评测，主要搜索引擎返回的相关结果的比率不足45%[3]，而且由于所采用机制、算法与适用范围等的不同，导致同一个检索请求在不同搜索引擎中的查询结果的重复率不足 34%[3]。元搜索引擎是建立在搜索引擎之上的，它以代理的角色，接受用户的查询请求，把查询请求转换成相应搜索引擎的查询表示，接受结果响应，以统一形式显示结果。元搜索引擎把用户的查询串分配给几个指定的成员搜索引擎，再将各成员搜索引擎所得结果分级排序，删去重复内容，然后给出查询结果[4]。

2 系统总体结构

元搜索引擎[5]主要由三部分组成(如图 1)：请求提交代理、检索接口代理、结果显示代理。从图 1 中可得知，元搜索引擎的技术重心在于查询前的处理(检索请求提交机制和检索接口代理)和结果的集成。结果集成的主要思想在于根据用户查询串与结果记录中的摘要信息、标题、问题关键词最小距离等多种信息计算相似度来对搜索结果进行排序。摘要和标题信息是根据每个成员搜索引擎检索页面 html 格式的不同分别从各个成员搜索引擎检索出的结果集中提取出[6]。

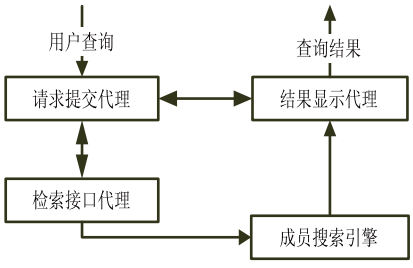

图 1 元搜索引擎原理图

选择百度、谷歌、雅虎、有道作为元搜索引擎的

① 收稿时间:2010-08-27;收到修改稿时间:2010-09-29

成员搜索引擎，各成员搜索引擎由结果满意值、查询转换、去除冗余网页地址、网页地址有效性、结果集成组成。

2.1 成员搜索引擎结果满意值

对于元搜索引擎来说，选择的搜索引擎越多，固然得到更全面的搜索结果，但是结果集成将花费大量的时间。为此本系统添加成员搜索引擎结果满意值功能。结果满意度对最终结果重新分配各个成员搜索引擎相应的比值。由于实际搜索引擎的性能和用户的请求是动态改变的，所以使用元搜索引擎的所有成员搜索引擎不利于系统整体性能的优化，通过添加该功能实现了用户与系统之间的交互，并提高了系统的整体效能。

2.2 查询转换

系统默认每个成员搜索引擎的查询数目 m=10，成员搜索引擎的个数 N=4，所以系统检索的数目范围为 10≤M≤40，保证系统的运行效率和信息覆盖面。

2.3 去除冗余网页地址

由于元搜索引擎是调用其它独立搜索引擎的引擎，亦称“搜索引擎之母”，就会出现多个独立搜索引擎进行检索时会搜索出相同的网页，同时，元搜索引擎的结果显示代理的作用之一就是负责将所有源搜索引擎检索结果的去重、合并[7]。使用两种方法对重复的 URL 进行去重处理：a.具有相同起始子串的 URL；b.URL 不同，但标题相同：设第 i 个摘要长度为 Li，第 j 个摘要长度为 Lj,如果 $L_i \geq 20$, $L_j \geq 20$，则与 length=20 个字符比较；如果 $L_i \leq 20$, $L_j \leq 20$，则选择 length=min(L_i, L_j)作为比较的字符个数，然后对两个摘要内容取长度为 length 字符串进行比较，一直将第 i 个摘要全查询完为止。

2.4 确定网页地址有效性

在对成员搜索引擎进行测试时发现有些搜索引擎的搜索结果无法正常显示网页内容，可能是在成员搜索引擎的索引库中仍存有其信息，但实际的网页已经被撤除，因此对于获得的超链结果集去除无效网页的地址。

2.5 元搜索 Agent 的结果集成

经过前 4 个步骤，系统获得了所选成员搜索引擎中有效、无重复信息的搜索结果，有利于计算每个结果相似度值，提高系统的查准率。

3 多信息融合相似度计算

本文提出了多信息融合相似度计算，包括基于摘要信息相似度计算、基于标题相似度计算和基于问题关键词最小距离相似度计算。

基于摘要信息的相似度计算是根据各页面的自动摘要信息中包含问题关键词的个数；基于标题的相似度计算是根据各页面对应的标题中是否全部含有问题关键词集；基于问题关键词最小距离的相似度计算是分别将问题关键词所处位置进行组合后，找到关键词最小距离。本文以“北京市的总面积是多少平方公里？”为例，说明多信息融合相似度计算。用户通过问题录入界面输入问题后，首先将问题通过问题分析模块进行处理，去除停用词后，获得问题类型、问题关键词，而后将数据送给元搜索的信息检索模块，该模块通过 4 个成员搜索引擎获得搜索结果集，并对其进行了去重、确定有效性等措施后，搜索结果的个数是 18 个，根据以下三种方法实现多信息融合相似度计算：

3.1 摘要相似度

设第 i 个摘要中第 j 个问题关键词的出现次数表示为 $keyWordsCount_{ij}$,$1 \leq i \leq N$, $1 \leq j \leq M$， $Max(\sum keyWordsCount_{ij})$表示在各摘要中问题关键词出现总次数的最大值，摘要与问题相似度为公式(1)：

$$abstractSim(Q,A_i) = \frac{\sum_{j=1}^{M} keyWordsCount_{ij}}{Max(\sum_{j=1}^{M} keyWordsCount_{ij})} \quad (1)$$

18 个结果集的摘要根据公式(1)获得基于摘要信息相似度值，通过 Excel 转换成图 2 柱状图。

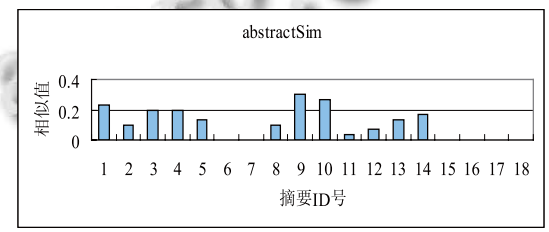

图 2　基于摘要信息相似度柱状图

3.2 标题相似度

在前期对各成员 SE 的测试过程中发现，大部分网页的标题是对整个页面内容的总括，考虑到该点，引入标题相似度。n_i 为第 i 个标题的统计变量，统计如果哪个问题关键词在标题中出现，则 n_i++，问题关键词个数为 realQL，所以标题与问题相似度为公式(2)：

$$titleSim(Q,T_i) = \frac{n_i}{realQL} \quad (2)$$

18 个结果集的摘要根据公式(2)获得基于标题相

似度值，通过 Excel 转换成图 3 柱状图。

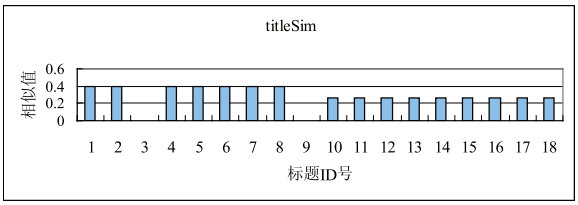

图 3 基于标题相似度柱状图

3.3 关键词之间最小距离

将关键词在摘要中出现的位置进行组合。例如，问题关键词集合中有 3 个关键词，分别用 a_1,a_2,a_3 表示，它们在摘要 i 中的位置：a_1 的位置(41、50、64)，a_1 出现 3 次；a_2 的位置(47、79)，a_2 出现 2 次；a_3 的位置(53、58、76)，a_3 出现 3 次，则共有 3*2*3=18 种组合，在这每种组合中选择最小位置和最大位置，以获得两者之间的距离。这样是为了解决在汉语问题中出现的词语所处位置不同，但问的却是相同的问题。

设 disAbstracti 表示问题关键词集全部在摘要中出现的最佳组合最小距离；minDis=Min(disAbstracti),1≤i≤N，则关键词最小距离为公式(3)：

$$disSim(Q,A_i)=\frac{1}{disAbsctract_i}\times \min Dis \tag{3}$$

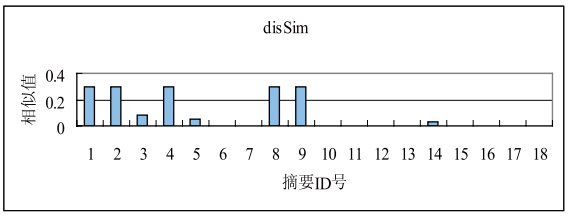

图 4 基于问题关键词最小距离相似度柱状图

18 个结果集的摘要根据公式(3)获得基于关键词之间最小距离相似度值，通过 Excel 转换成图 4 柱状图。

3.4 搜索获得的摘要结果集和问题总的相似度：

公式(4)：

$$\begin{aligned} Sim(Q,Abstract_i)=&\alpha\times absctractSim(Q,A_i)\\ &+\beta\times titleSim(Q,A_i)\\ &+\gamma\times disSim(Q,A_i)\end{aligned} \tag{4}$$

其中 α + β + γ =1，显然 0≤Sim(Q，A_i)≤1。

摘要和标题是直接反映网页内容的标识，并根据系统测试 α =0.35，β =0.4，γ =0.25。

18 个结果集的摘要根据公式(4)获得摘要结果集和问题总相似度值，通过 Excel 转换成图 5 柱状图。

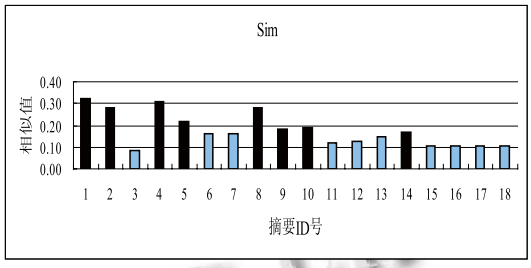

图 5 相似度柱状图

根据公式(4)获得各自总的相似度，从中挑选出相似度最高的 8 个搜索结果按照由高到低的顺序作为后续答案提取模块的依据。

4 结语

通过元搜索引擎提高了搜索结果的覆盖面，利用多信息融合相似度计算提高了搜索结果的查准率。但考虑到问题分析模块可以通过语义分析获得与用户需求相关信息，信息检索模块在多信息融合相似度计算基础上引入语义分析，将能进一步提高查准率，为答案提取模块提供广泛而全面的候选结果集。

参考文献

1 李晓明,闫宏飞,王继民.搜索引擎——原理、技术与系统.北京:科学出版社,2005.

2 徐宝文,张卫丰.搜索引擎与信息获取技术.北京:清华大学出版社,2003.

3 李广建,黄昆.元搜索引擎及其主要技术.情报科学,2002,2(2):22-27.

4 王忠,程磊.基于元搜索引擎的个性化 Web 信息采集.计算机工程与设计,2009,30(13):3117-3119.

5 曹林,韩立新,吴胜利.元搜索引擎排序技术综述.计算机应用研究,2009,26(2):411-414.

6 王忠,程磊.基于元搜索引擎的个性化 Web 信息采集.计算机工程与设计,2009,30(13):3117-3119.

7 谢蕙,秦杰.基于元搜索的网页消重方法研究.计算机系统应用,2008,17(8):94-96.