●王兰成　李　超

改进的中文同义词相似匹配方法

摘　要　信息检索的核心技术是文档集与提问集的相似匹配。目前基于关键词的字面匹配方法和基于词义的概念匹配方法各有优势与不足。在数字图书馆文献检索中可以综合两者的优势。设计一种改进的中文同义词相似匹配方法较现有两种方法优越,并具有更好的应用性。图 2。表 3。参考文献 2。

关键词　中文同义词　信息检索　自然语言处理　控制词表
分类号　G252.7

ABSTRACT　The core technique for information retrieval is the similarity matching of file set and query set. There are differences between keyword-based word matching method and meaning-based concept matching method. In the development of document retrieval for digital library, we can utilize the advantages of both methods and design a new, improved Chinese synonym similarity matching method. 2 figs. 3 tabs. 2 refs.

KEY WORDS　Chinese synonym. Information processing. Natural language processing. Control thesaurus.
CLASS NUMBER　G252.7

数字图书馆的文献信息检索已经成为人们关注的热门技术之一。目前的信息检索主要是基于关键词检索方法,网上检索几乎都采用基于关键词匹配的全文检索技术。关键词检索虽然相对成熟,但在表达检索者的意图上还存在一定的缺陷。随着计算机技术和人工智能的发展,自然语言越来越多地应用于信息检索领域,它涵义丰富,对概念表达灵活,这为数字图书馆建设提出了更多的研究课题。

1　中文同义词的相似匹配方法

同义词在语言学、情报学中都存在,但其含义并不相同。汉语词典中将同义词定义为"词义完全相同或相近的词",如"自行车"和"单车","土豆"和"马铃薯"等等,这是严格意义上的同义词。情报学中同义词的概念相对比较宽泛,情报检索语言中叙词间的语义关系主要有同义关系、属分关系和相关关系 3 种。同义关系是指几个同义词或准同义词之间的一种语义关系,包括词义完全一致的两个表达同一概念的真同义词、表达同一概念语义相近的两个近义词、繁称词和简称词、不同的译名、学名与俗名词等情形;属分关系是指具有上位概念(属概念)和下位概念(分概念)的叙词之间的语义关系,包括属种关系、整体部分关系以及包含关系 3 种情形;相关关系是指除同义关系和属分关系以外的语义关系,这些关系包括有部分重合的两个概念之间的交叉关系、外延总和等于其上位概念全部外延的两个并列概念之间的矛盾关系、外延总和小于上位概念的两个并列且相互对立的概念之间的对立关系、同一个属概念之间的并列关系。目前,自动识别中文同义词的方法主要有以单汉字(即语素)为单位的字面相似匹配方法和以词概念(词素)为单位的语义相似匹配方法。

1.1　以语素为单位的字面相似匹配方法

汉语构词法有两个重要特征:(1)意义相同或相近的语词大多包含有相同的字,即情报学意义上的同义词多数表现出字面相似的特点。如"英文编目"与"西文编目"、"军队院校"与"军事院校"等,这就为通过计算两个语词的字面相似程度来判断它们在语义上的相似程度提供了一定的理论依据。(2)复合词往往中心词在后,词素层层限定。如"微型电子计算机"、"可擦除可编程只读存储器"、"山村小学教师"等。这一构词特点也称为汉语构词的"重心后移"现象[1],表明复合词中每个单词素的含义对整个词的意义的影响是不一样的,越靠后的词素,其含义对整个词的意义影响越大。这又为计算复合词的相似度并在考虑字面相似度同时也考虑词素在复合词结构中的位置赋予了相应的权值并参与相似度计算。

设参与匹配的语词 A 含有的单字字数为 n(a),

语词B含有的单字字数为n(b),两者含有的相同字的个数为n(ab),两词的相似度为P(a,b),则最简单的字面相似度计算方法可以表示为:

$$P(a,b) = \frac{n(ab)/n(a) + n(ab)/n(b)}{2} \times 100\% \quad (式1)$$

考虑词素权重可以得出另一种算法。假定字数为i的语词从左至右每一个单汉字的权值分别为Q(1)=1,Q(2)=2,…,Q(i)=i,则算法可以表示为:

$$P(a,b) = p \times \frac{n(ab)/n(a) + n(ab)/n(b)}{2} + q \times L(a,b) \times \frac{\sum Q[n(ab)]/\sum Q[n(a)] + \sum Q[n(ab)]/\sum Q[n(b)]}{2} \quad (式2)$$

其中,p、q分别表示匹配字数和词汇结构对相似度的影响系数,p,q≤(0,1)且p+q=1。为了和文献[1]有可比性,这里仍取p=0.6,q=0.4。

L(a,b)是参与匹配的两个词的长度比值,其取值作如下约定:

当n(a)<n(b)时,L(a,b)=n(a)/n(b);当n(a)≥n(b)时,L(a,b)=n(b)/n(a)。

下面举例说明。

例1 选取"军事情报管理"作为语词A,"军事信息管理"作为语词B,计算它们的相似度。

$$n(a)=6, n(b)=6, n(ab)=4, L(a,b)=6/6=1,$$

$$P(a,b) = 0.6 \times \frac{4/6+4/6}{2} + 0.4 \times 1 \times \left(\frac{1+2+5+6}{1+2+3+4+5+6} + \frac{1+2+5+6}{1+2+3+4+5+6}\right) \times \frac{1}{2} = 66.7\% \quad (式3)$$

例2 选取"中华人民共和国"与"中国"两词参与匹配计算,计算它们的相似度。

$$n(a)=7, n(b)=2, n(ab)=2, L(a,b)=2/7,$$

$$P(a,b) = 0.6 \times \frac{2/7+2/2}{2} + 0.4 \times \frac{2}{7} \times \left(\frac{1+7}{1+2+3+4+5+6+7} + \frac{1+2}{1+2}\right) \times \frac{1}{2} = 45.9\% \quad (式4)$$

例3 选取"土豆"与"马铃薯"两词。容易得出,P(a,b)=0。

显然,例1得出的结果基本符合要求,例2有比较大的误差,例3算出的结果可以说是一个错误。

1.2 以词素为单位的语义相似匹配方法

以语素为单位的字面相似匹配方法,尽管考虑了汉语构词的重心后移特点,但从概念的角度来看,汉语通常是以词素,而不是单汉字为语义单位的。因此,根据语义进行匹配,同时以词素为单位赋予权值更加符合语言的本质特点,这就是语义相似匹配方法的基本思想[2]。方法是首先编制一部语义精良的同义词表,然后将查询式采用自动分词的技术切分为一个个词素,在进行同义词相似匹配时,只需根据词表中收录的同义词词条,结合权重计算,进行匹配。构造如表1结构的同义词表。

表1 结构同义词表

记录号	主题词	同义词
1	中央军委	中央军事委员会;中共中央军委;军委(中央)
2	秘密组织	地下组织
3	番茄	西红柿
4	自行车	脚踏车;单车
…	…	…

这种词表实际上能对用户输入的查询式起到规范和控制作用,故也称后控制词表。当我们用这个词表进行语义检索时,首先对复合词以及句子进行分词处理,得到单个词素,对词素从语义层面进行相似匹配,然后仍然考虑词组或句子的结构关系,结合权值进行计算,得到结果。

例4 对"军事情报管理"与"军事信息管理"进行词素相似匹配计算。

"军事情报管理"分为"军事"、"情报"与"管理"3个词,"军事信息管理"同样分为"军事"、"信息"、"管理"3个词。由于词表中收录有"信息"与"情报"的同义关系记录,因而计算结果为:P(a,b)=100%。

例5 对"中央军委"与"中央军事委员会"进行词素匹配计算。

词表中有相关记录,计算结果为:P(a,b)=100%。

例6 对"电子邮件"与"伊妹儿"进行词素匹配。

词表中无相关记录,计算结果为:P(a,b)=100%。

2 改进的中文同义词相似匹配方法

基于语素的字面匹配以单个汉字作为匹配基本

单位而避免了分词的障碍，无需词表故不会出现遇到新词无法进行匹配计算的情形。但在汉语语词中，有的在字面上相似但语义上却相差甚远，有的尽管字面上不相似而表达的却是相近的甚至同一个概念。因此，单纯的字面匹配方法检索结果往往偏离用户的真实检索意图，误检率较高；如果用户输入的查询式稍有偏差，检索系统就无法确定用户的真正需要，因而无法将那些实际上是用户需要的但在字面上却与查询式不相似的记录提交给用户，造成漏检。以词素为单位的语义相似匹配方法从概念层面进行相似匹配，遵循了汉语本身的规律，匹配结果较之字面匹配方法要理想得多。采用这种方法的前提是要编制一部词表。编制词表是一项浩大的工程，词表在编制时只可能收集有限的词素，无法应对新词的不断出现以及不断扩充的同义词规则。当匹配词在词表中尚未收录时，词表法往往显得无能为力，如例6就暴露出它的一些不足：由于词表中尚未收录“电子邮件”的同义词相关记录，匹配结果出现不可接受的偏差。

综合考察上述两种计算同义词相似度的方法，不难发现，两者各有利弊：以语素为单位的字面相似匹配从理论上讲可以对任意两个语词进行计算，但其结果的精度难以保证；以词素为单位的语义相似匹配在可以保持高精度的同时，对尚未收录的新词的相似计算却很困难。若把两者结合起来，不失为一种好的方案。这种方案的总体流程如图1所示。

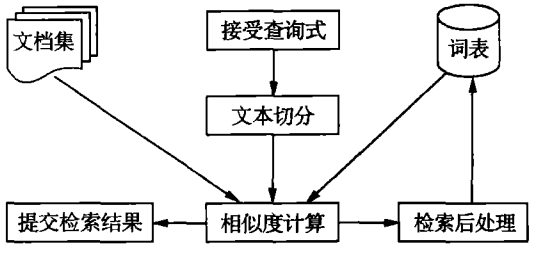

图1　改进的同义词相似匹配流程

与前面两种相似匹配方法相比，改进的中文同义词相似匹配方案主要有两点改进：一是词表结构的改进。在信息检索系统中，存储与检索效率存在着得失互换的规律，即在建立数据库和索引时做的工作越多，检索系统的效率越高，检索越方便易行；反之，如果在存储这一环节只做少量工作（例如，只准备粗糙的后控词表和简单的索引），检索时就会比较麻烦和复杂并且难以达到较高的检索效率。如果设计词表时把概念逻辑中的另外两种关系也考虑进去，使整个词表记录构成一个有机整体，将大大提高检索的方便性。同时，通过设计新的算法，便可以实现同义词扩展检索和相关词参照检索。我们设计的改进词表结构如表2所示。二是算法的改进。用户检索信息有如下特点：第一，能提供与所需要的文档含义相同或相近的检索式。如用户要查找与“微型计算机组装”有关的资料，输入的查询式不外乎“微型电脑组装知识”、“微机装机资料”、“电脑组装技术”，或是“电脑技术”、“电脑知识”这种上位概念。第二，在表达检索要求时，倾向于使用自然语言甚至口语来表达相关概念和构造查询式。事实上也正是由于这一特点，使同义词相似匹配研究变得非常必要。如用“招飞”来表达“招收飞行员”，用“梯恩梯”来表达“三硝基甲苯”等。此外还有其他特点，如用户构造的查询式在表达概念时仍具有“重心后移”特点，等等。

表2　改进的词表结构

记录号	主题词	同义词	上位词	下位词	相关词
1	情报检索	情报查找；信息查找…	3	2	5
2	计算机信息检索	文献计算机检索…	1	0	4
3	信息技术	IT；IT技术	0	1	0
4	手工信息检索	手工检索；人工信息检索…	1	0	2
5	信息获取	消息获取；情报收集…	0	1	1
…	…	…	…	…	…

注：表中“上位词”、“下位词”以及“相关词”字段中的数字表示记录号。

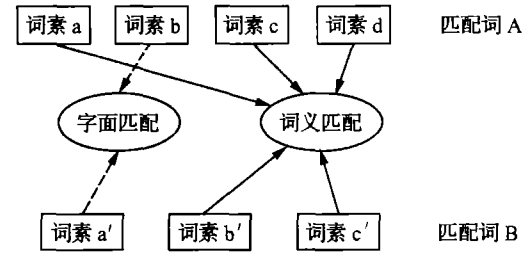

图2　改进的同义词相似匹配算法示意

结合前两种匹配方法，设计新算法的思路如图2所示。假定匹配词A与B通过分词技术之后分别得

到词素 a,b,c,d 和词素 a',b',c'。这些词素中,有些可能是词表中没有收录的词,如词素 b,a',另一部分是词表中已收录的词素,如词素 a,c,d 以及 b',c'。在设计算法时,先考虑语义匹配,然后对无法采用语义匹配方法的新概念采用字面匹配的方法。

3 新方法的测试与结果分析

为了检验新方案的可行性和它较之前两种方法的优越性,我们设计了一个同义词识别实验系统。测试过程中分别选择了几组单纯词、复合词、短语和短句进行比较。具体测试结果如表3。

表3 测试结果

匹配词 A	匹配词 B	方法1 P(a,b)	方法2 P(a,b)	方法3 P(a,b)
计算机模拟	计算机仿真	0.52	1	1
中国	中华人民共和国	0.46	1	1
因特网	国际互联网	0.26	1	1
下海	经商	0	1	1
中国设备进口	国家技术设备引进	0.53	0.67	0.83
军队信息管理学	军事情报学	0.30	0.72	0.79
非典型肺炎预防治疗工作	非典防治工作	0.69	0.83	0.94
电子邮件	伊妹儿	0	0	0

注:方法1为字面匹配,方法2为语义匹配,方法3为改进匹配。

从测试情况看,新方案较之纯粹的字面匹配或语义匹配方法确实有许多改进,计算精度总体上有较大提升。在选取的8组词中,前4组是单个词素,且在词表中收录了相关的同义词记录,因而除了方法1之外,方法2、方法3都取得了理想的匹配结果。而第5、6、7组是复合词,第5组中,只有"设备进口"与"技术设备引进"是同义词;第6组中只有"信息"与"情报"为同义词;第7组中只有"预防治疗"与"防治"为同义词。在对5、6、7这三组的相似匹配中,方法3的优势充分体现出来:三组匹配结果中,方法3得出的精度是最高的。而对于最后一组"电子邮件"与"伊妹儿",由于两者在字面上缺乏相似性,并且词表中尚未收录电子邮件的相关同义词记录,故匹配的结果还是不尽如人意。这说明,尽管方法3在前两种方法上有了不少改进,但仍然需要进一步改进。

4 进一步的工作

改进的方法在保证比较理想效果的同时,又能对任何两个语词进行比较,较之单纯的字面或语义匹配算法有了较大改进。该方案主要在后控制词表的结构和相似匹配算法上做了一些改良,实验结果证实了新方案的可行性与优势。然而在实际应用中,涉及的新词将不断增加,但词表不变或更新慢,其匹配功能将越来越弱。如果加入检索后处理过程,则可望达到更加理想的匹配精度。具体过程可描述为:①新建一个自由词词表,用于保存在匹配过程中遇到的新词;②识别并获得新词,并将新的记录写入自由词表;③通过词频统计方法,对自由词表中收集的词素进行分辨和筛选,即通过对新词在匹配中出现的频率,来确定是否应该将其归入同义词表;④将筛选出的符合要求的新词素添加到同义词表中,更新词表;⑤再进入下一次的检索过程以及检索后处理过程。

用自然语言实现概念检索是一项复杂艰巨的工程,而同义词概念的相似匹配,正是概念检索要研究和解决的问题之一。随着研究的不断深入和人工智能技术的不断进展,计算机自动识别汉语同义词的研究必将取得新的进展,概念检索技术也将会取得新的突破。

参考文献

1 王源等. 后控规范的计算机处理. 现代图书情报技术, 1993(2)

2 朱毅华,侯汉清,沙印亭. 计算机识别汉语同义词的两种算法比较和测评. 中国图书馆学报,2002(4)

王兰成 解放军南京政治学院上海分院信息管理系教授,博士生导师。通信地址:上海市南京政治学院上海分院信息管理系。邮编200433。

李 超 南京政治学院上海分院信息管理系硕士研究生。通信地址同上。 (来稿时间:2004-08-25)