·2007年年会征文选登·

自动标引通用评价模型研究*

章成志[1,2]　周冬敏[3]　苏新宁[3]

(1. 南京理工大学信息管理系　南京 210094)
(2. 中国科学技术信息研究所　北京 100038)
(3. 南京大学信息管理系　南京 210093)

摘　要　自动标引的评价是度量一个标引系统的性能的重要手段。针对常规自动标引评价方法存在的评价结果不能完全反映真实标引结果,以及评价成本高的情况,本文提出一种通用的自动标引评价模型,该模型有效利用外部资源,根据有参照情况与无参照情况,分别对标引结果进行评价。实验结果表明,自动标引通用评价模型能增加评价的可靠性并降低评价的成本。

关键词　自动标引　评价模型　语义相似度　相似度计算

1　引言

自动标引包括关键词自动提取(又称自动抽词标引)与自动赋词标引两种类型。关键词自动提取是一种识别有意义且具有代表性片段或词汇的自动化技术[1]。由于关键词是表达文件主题意义的最小单位,因此大部分对非结构化文件的自动处理,如自动摘要、自动分类与聚类、事件检测与跟踪、概念检索关联知识分析、自动问答等,都必须先进行关键词提取的动作,再进行其他的处理。可以说,关键词提取是所有文件自动处理的基础与核心技术[2]。

目前大多文档都不具有关键词,同时手工标引费力费时且主观性较强,因此关键词自动标引是一项值得研究的技术[2]。由此引发的标引结果有效评价问题也成为一个亟需解决的问题,而评估关键词自动抽取的性能并不是一件容易的事情[3]。

自动标引评价方法可以分为两种,即内部评价方法与外部评价方法。内部评价方法是从标引系统的目标出发,具体方法包括标引结果比较法、要点评价法、用户可接受性评价[4]。内部评价方法主观性高,存在评价不一致性情况。同时,这种往往忽视关键词之间存在的语义关联问题,造成评估结果的不可靠性。为了克服内部评价方法存在的缺点,研究人员提出外部评价方法。外部评价方法则从标引所提供的功能出发[5],将标引结果应用于特定的任务,如文本分类[6]、聚类或检索[7],根据标引系统对该任务的促进作用来评价标引系统的性能。外部评价方法是针对特定任务,任务中所使用语料质量、规模对评估结果有很大影响。同时,特定任务本身的计算量通常会超过自动标引过程本身的计算量,这样使得评估的速度难以满足实际需要。因此,外部评价方法一般较少被用于评价自动标引性能。

本文针对上述问题,提出一种通用的自动标引评价模型,该模型有效利用外部资源,根据有参照情况与无参照情况,分别对标引结果进行评价。实验结果表明,自动标引通用评价模型能增加评价的可靠性,评估速度快,并降低评价的成本。

2　一种通用的自动标引评价模型

* 本研究受"十一五"国家科技支撑计划重点项目(2006BAH03B00)、2006年江苏省研究生培养创新工程项目资助。

2.1 自动标引通用评价模型的引入

对于自动标引系统来说,已有的评价标准是否能很好地反映标引性能已成为人们关注的问题。本文将自动标引结果的评价分为两种:有参照情况下的(有评价参照的关键词集合,一般为已标注的关键词集合)自动标引结果的评价,无参照情况下(没有评价参考的关键词集)的自动标引结果的评价。无论是采用有参照还是无参照方法,评价的主要依据都是考察待评价的标引关键词集合反映文章的主题内容的程度。在有参照的情况下,将原文已有的关键词作为该文本主题内容的"替代者",这里称之为标准关键词。然后将待评价的关键词集合和标准关键词集合进行相似匹配,匹配程度反映了待评价关键词表达文章主题的程度。本文利用相似度方法来进行匹配程度计算。

在无参照的情况下,评价思路为先获得文本内容的"替代者",再将"替代者"作为参照对待评估关键词进行评价。获得原文的"替代者"的基本方法有两种。第一种方法是找到一个最优的标准的关键词集合,以它作为参照来评价,第二种方法是对文章主旨内容直接进行内容分析,获得原文的概念向量,然后将待评价的关键词集合和原文概念向量进行相似度计算。实际上,最能代表文本主题的就是文章本身,关键词总会割裂原文的语义[8],而且,寻找最优的标准关键词集合正是自动标引要研究的内容,在进行自动标引评价时,我们事先无法得到最优的标准关键词集合,但能确定原文的一个概念向量,因此,我们采取原文概念向量与待评估关键词集相似度评估的方法。原文的概念向量(Concept Vector,CV)能通过关键词的概念转换或者潜在语义索引(Latent Semantic Index,LSI)[9]得到。本文使用简化的方法,即直接用向量空间模型(VSM)[10]表示文本,以VSM表示的向量作为文本概念向量。

下面给出自动标引通用评价模型的总体框架,并描述评价的基本原理和关键技术。

2.2 自动标引通用评价模型框架

根据上文分析,本节给出自动标引通用评价模型的框架图如图1所示。通用评价模型的输入为文献集合以及待评价的每篇文档的标引关键词集合。另外,文档集合中的文档可能自身就带有人工标注的关键词集合(在某些情况下,人工标引的关键词集合并不理想,不能用来作为标引评价中的参照),某些文档也可能没有经人工标引的关键词。

通用评价模型的评价部分要解决的关键问题即为:待评价的关键词表达文档主题的程度的度量。我们通过待评价的标引关键词集合与参照集合之间的相似度计算来解决这个问题。

通用评价模型的输出部分为模型结果的可视化显示,包括经过相似度计算过后的各种标引模型的标引结果的查准率(P)、召回率(R)、F1值等的图形化显示等。

2.3 通用评价模型中的评价方法描述

下面描述在有参照和在无参照情况下的标引结果评价方法。

2.3.1 有参照时的标引结果评价方法

(1) 基本原理

有参照的关键词评价方法主要是利用相似度原理对传统的评价方法进行改进。如表1与表2所示,传统的标引评价指标,如P、R、F_1值等是以自动标引与人工标引产生的关键词之间的精确匹配(Exact Match)为基础的。这种评价方法往往忽视自动标引关键词集合和参照关键词集合之间存在的语义关联问题,即两关键词集合中的关键词之间存在同义词或部分匹配情况。当标引结果中存在大量由赋词标引方法得到的标引词时,这种评价方法对评价结果的负面影响更加剧烈。例如,针对一篇主题为"土豆栽培"的文章,原文标引的关键词可能经过赋词标引得到,即标引结果包含经主题词表规范的词语,如原文标引词为"马铃薯",自动标引结果可能给出的结果包含关键词"土豆"。在传统评价方法中,"土豆"与"马铃薯"为不同的关

键词，导致评价结果 P、R、F1 值不能反映自动标引真实性能。

有参照情况下，通用评价模型评价方法的基本思想是：考虑待评价标引关键词集合与参照标引关键词集合之间存在的语义相似性，并借助外部资源对相似程度进行有效的量度，从而给出更合理的结果评价。续上例，经相似度计算后，评价系统认为“土豆”与“马铃薯”为同义词，即机器标引为“土豆”是正确的。因此，在传统评价方法的基础上引入待评价与参照关键词集合间相似度计算方法，将对自动标引评价结果起到积极作用。

为了更好地说明该问题，图 2 给出有参照情况下的标引结果评价基本原理的示意图。图 2(a)描述传统评价方法，只考虑到待评价标引关键词集合与参照的标引关键词集合之间的关键词精确匹配情况，即图中的交叉部分，记为 a。采用相似度改进后，如图 2(b)所示，集合 b 部分与集合 c 部分存在 k 对语义上较相似的关键词对，可依据这些关键词对的相似度计算结果，建立两者之间的语义关联。Δc 为采用相似度改进方法后，同义词匹配或者相似度高于给定阈值的关键词匹配的情况，$a+\Delta c$ 为采用相似度改进方法后，待评价的标引关键词集合与参照的标引关键词集合之间的关键词匹配的总体情况。

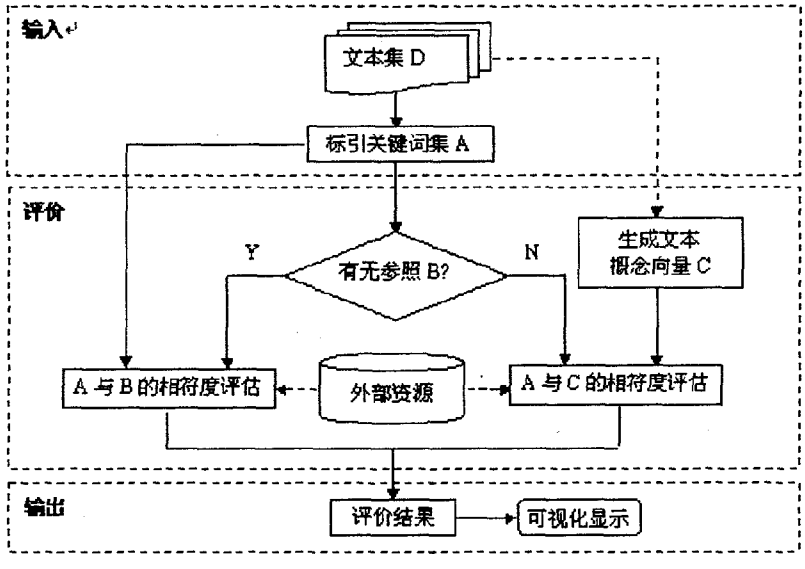

图 1　自动标引通用评价模型框架图

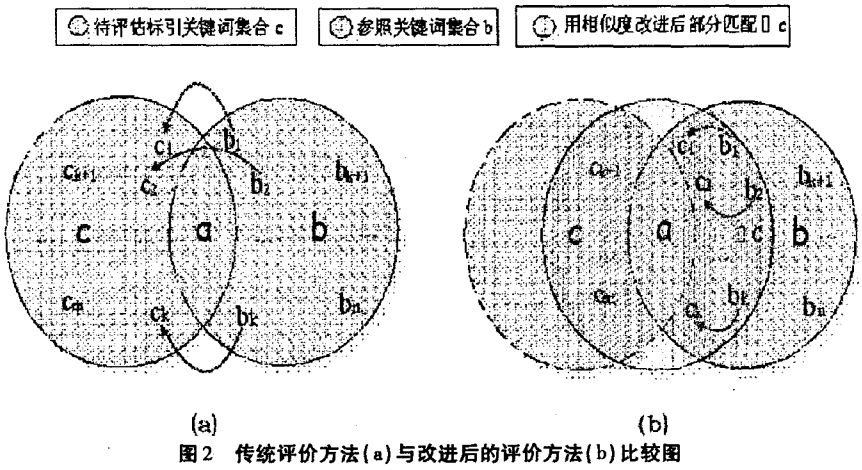

图 2　传统评价方法(a)与改进后的评价方法(b)比较图

假设测试集中词语总数为 n,则自动标引系统标引结果如表 1 所示。将人工标引的结果分为两种情况,分别为人工标引为关键词的情况(即(a + c))与人工标引为非关键词的情况(即(b + d))。人工标引为非关键词的情况,就是将人工标引关键词后,文本剩下的词作为非关键词。同理,自动标引结果也可以分为这两种情况,其中,(a + b)为标引系统标引的关键词总数。

表 1 标引结果评价列联表

	人工标引为关键词	人工标引为非关键词
标引系统标引为关键词	a	b
标引系统标引为非关键词	c	d

表 2 标引评价指标对照表

评价指标	传统方法	通用评价模型改进后
查准率(P)	$P=\frac{a}{a+b}$	$P'=\frac{a+\triangle c}{a+b}$
召回率(R)	$R=\frac{a}{a+c}$	$R'=\frac{a+\triangle c}{a+c}$
F1 值	$F_1(P,R)=\frac{2PR}{P+R}$	$F_1'(P',R')=\frac{2P'R'}{P'+R'}$

考虑到待评价的标引关键词集合与参照的标引关键词集合之间存在的语义关系后,自动标引结果的查准率 P、查全率 R、F_1 值计算方法修正分别为 P′、R′、F_1′,如表 2 所示。

(2)评价步骤

如图 1 所示,有参照关键词集合评价步骤包括四个步骤:①选取待评价标引关键词集合与参照关键词集合。②计算集合相似度。利用主题词典、语义词典或者本体等外部资源计算参照关键词集合和待评价关键词集合的相似度。③标引评价指标计算。经过相似度计算后,分别计算得到新的 P、R、F_1 值的,记为 P′、R′、F_1′。④评价结果的图形化显示。

(3)$\triangle c$ 获取算法描述

$\triangle c$ 为待评价的标引关键词集合与参照的标引关键词集合之间存在的同义词与语义相似部分。$\triangle c$ 的获取算法描述如图 3 所示。

算法:$\triangle c$ 获取算法描述

输入:自动标引关键集合排除精确匹配后的关键词集合,记为 $b=\{b_1,b_2,\ldots\ldots b_n\}$;

参照的标引关键词集合排除精确匹配后的关键词集合,记为 $c=\{c_1,c_2,\ldots\ldots c_m\}$;

关键词相似度阈值 η;

输出:$\triangle c$

步骤:Step1:对两个集合中每对关键词进行相似度计算,得到关键词相似度矩阵 M,$M=\{t_{ij},i\in[1,m],j\in[1,n]\}$,$t_{ij}$ 表示集合 c 中关键词 c_i 与集合 b 中关键词 b_j 的相似度;

Step2:If $t_{ij}>\eta$ then

关键词 c_i 与中关键词 b_j 相似(特别的,若 $t_{ij}=1$,则认为这两个关键词为同义词)

Else

关键词 c_i 与中关键词 b_j 不相似;

Step3:遍历矩阵 M,对每一个元素 t_{ij},若 $t_{ij}<\eta$,则将 t_{ij} 赋值为 $t_{ij}=0$;

Step4:遍历矩阵 M,找出矩阵 M 中值最大的元素 t_{ij},将 t_{ij} 的值取出,记为 (i,j,t_{ij}),将第 i 行和第 j 列所有元素(包括 t_{ij})的值都设为 0,原矩阵变为 M';

Step5:对矩阵 M' 重复 Step4,直到矩阵中的所有元素的值均为 0,得到的三元组为:$\{(i,j,t_{ij})\}$;

Step6:根据 $\triangle c=\sum t_{ij}$(t_{ij} 为筛选出来的关键词对的相似度值)求得 $\triangle c$。

图 3 获取算法

值得注意的是,参照关键词集合作为原文的替代物,其概括文献内容的能力直接关系到评价结果的好坏。因此,有必要对“参照”关键词进行预先检验。检验的方法可以采用无参照情况下标引关键词的评估方法。

2.3.2 无参照时的标引结果评价方法

（1）基本原理

如图1所示，在没有参照的情况下，通用评价模型试图找到原文的一个参照，作为原文内容的“替代者”。该过程类似GF自动标引的前期过程，即在词频统计的基础上，考虑词语的tf * idf值、词语首次出现的位置、词语出现的重要位置信息[11]等，最后得到一个带有关键词权重的词语向量。与之不同的是，这个向量的关键词数量并没有严格的限制，以更大程度地反映原文内容为标准。接着将这个“虚拟”参照与待评价的标引关键词集合进行相似度计算，以相似度计算结果来反映自动标引的关键词表达文本主题的能力。

（2）评价步骤

如图1所示，在无参照的情况下，关键词集合评价步骤包括四个步骤：①生成文本概念向量。②计算集合相似度，即计算待评价的关键词集合与由步骤①得到的关键词集合相比较。其本质在于比较一个主题很浓缩的“替代者”和一个主题较浓缩的“替代者”之间的匹配程度。我们利用从两个集合之间的相似程度来计算这两个“替代者”的匹配程度。具体计算方法参见2.4小节。③标引结果评分。根据步骤②的相似度计算结果给出评价。根据相似度的大小来给标引结果赋一个分值，以此表明标引结果的好坏程度。

需要指出的是，由于没有参照关键词集，可采用外部评价法作为辅助，即利用自动标引结果用于文本分类、聚类或检索，通过这些应用的结果评估自动标引系统的性能[6-7]。

2.4 通用评价模型中的关键技术——相似度计算

计算关键词之间或关键词集合之间的相似度是自动标引通用评价模型的中的关键问题。

值得注意的是，这里的相似度计算是指在排除精确匹配后才进行的。本节从两个方面说明相似度计算方法。

2.4.1 两个关键词之间的相似度计算

两个词语之间的相似度可以通过词形或语义方法度量。词形相似度也即词语字面相似度。当两个词语之间的词形相似度大于某个阈值时，就认为它们是相似词。编辑距离算法[12]是一种比较流行的基于词形相似度计算方法。语义相似度算法可以克服词形相似度算法中以字为单位，而不考虑语义的缺点。对于两个词语间的语义相似度，根据所利用外部资源的不同，计算方法也有所不同。如果是主题词表，则借鉴O. Medelyan在计算标引词之间的相似度时采用的方法，可以根据词语在主题词表中的用、代、属、分、族、参等关系进行计算[13]。如果所利用资源是语义词典，可以考虑词语间的层次结构，如果是本体，可以考虑更多的属性和概念之间的相似关系。将不同资源下的相似度计算结果进行集成，有助于克服片面性，从整体的角度来考虑相似度。在进行词语的相似度计算时，除了考虑利用主题词典、语义词典或本体等“静态”的外部资源之外，还可以利用大规模语料库进行词语的相关度度量。1999年Dagan等依据于语料库，使用概率模型来计算词语的距离[14]。

本文在进行词语的相似度计算时，主要利用基于语义词典的相似度计算方法，并以词形相似度计算方法作为辅助，具体计算方法可参见文献[15]。

2.4.2 关键词集合中关键词之间的相似度计算

在关键词相似度计算的基础上，可以进行关键词集合中关键词之间的相似度计算。

（1）有参照时的关键词间的相似度计算

该计算方法考虑的情况是有参照关键词情况下，待评价的标引关键词集合（记为A）和参照的标引关键词集合（记为B）在数量上大致相当的情况。假设A和B为元素数量相当的关键词集合，对于A中的每一个元素，依据词语相似度计算法，在B中找到一个和它相似度最高的元素和其进行对应。相似度计算结果表明每个词对之间的语义相似程度。

(2) 无参照时的关键词间的相似度计算

不同于有参照情况下的关键词集合之间的相似度计算方法,在无参照时,通常文本的"替代者"(本文使用文本的概念向量表示)通常为一个待评价的标引关键词集合大得多的词语集合,并且集合内每个元素都带有表示该词语重要性的权值信息。假设两个集合A和B。其中A为原文的概念向量,表示为一个较大的关键词集合形式,且每个元素都附带有权重信息。B为待评价的标引关键词集合。对于A中的任意一个元素,求出它与B中每一个元素之间的相似度,取相似度最大值作为A与B中关键词之间的相似度。

3 评价模型的应用与性能分析

3.1 自动标引通用评价模型的应用

本节简要说明自动标引通用评价模型应用于各种自动标引模型的性能评估的情况。本文以人大报刊复印资料[16]"人大2005年一季度经济类专题"库中的经济类论文600篇作为数据集,以6个标引模型为评价对象,给出了基于通用评价模型的评价结果。其中,数据集中每篇文献的原文中均有2-7个数目不等的关键词。在进行有参照的标引结果评价时,我们直接将这些已经标引好的关键词集合作为参照。在进行无参照的标引结果评价时,我们则"隐藏"这些标引结果,以原文的概念向量作为"虚拟"参照。

在通用评价模型中,本文利用《同义词词林》作为语义词典,进行关键词集合之间的语义相似度计算,将词形相似度方法作为度量关键词集合的相似度的辅助方法。

3.1.1 有参照情况下的评价

有参照情况下的评价目的在于对通用评价模型有参照情况下的评价方法进行性能分析。给定带有关键词的文章,并对文章分别采用不同的自动标引方法进行标引,得到的几组标引关键词集合,然后使用通用评价模型分别对其进行评价。通过通用评价模型的评价结果与真实情况的接近程度来说明有参照情况下通用评价模型的性能。

封闭测试集中,文章总数为500,其中关键词总数为1821,平均每篇文章的关键词数为3.64。赋词标引词数为209,赋词标引比率为11.48%。开放测试集中,文章总数为100,其中关键词总数为354,平均每篇文章的关键词数为3.54。赋词标引词数为41,赋词标引比率为11.58%。

使用传统评价方法时的评价结果显示如图4所示。使用通用评价模型进行评价后,评价结果如图5所示。由图4、图5可以看出,在使用通用评价模型对6个标引模型(分别为:MLR、Logit、SVM、CRF以及两个基准模型,由于篇幅所限,该部分内容将另文介绍)进行重新评价后,评价的结果发生了一定程度的变化。根据我们对评价结果的分析发现,经过相似度计算后,参照集合与自动标引结果集合中的部分关键词进行了语义的匹配,其中有部分同义词关系被识别出来。基于通用评价模型的评价结果更接近真实评价结果。

值得注意的是,有参照情况下标引结果评价方法的可信度强烈依赖于参照的质量。事先借助无参照情况下评价方法对参照本身进行预评估,将有参照和无参照的评价方法结合起来进行评价是我们今后需要进一步研究的工作。

3.1.2 无参照情况下的评价

该部分评价的目的在于对通用评价模型无参照情况下的评价方法进行性能分析。给定一组文章,并对文章分别采用不同的自动标引方法进行关键词标引,得到几组标引关键词集合,然后使用自动标引通用评价模型分别对其无参照情况下的评价。通过模型得分情况比较模型的优劣程度。

无参照情况下,本文直接利用VSM来表示文本的概念向量,并且只取权值排在前20位的词语一起构成原文的"虚拟"参照。计算标引结果组成的关键词集合与原文的"虚拟"参照之间的相似度,并依据表3进行结果打分。本文利用600篇文献作为评价对象,无参照情况下的自动标引评价结果如图

6所示。由图6可以看出，在给定的数据集下，无参照情况下的自动标引评价结果与有参照情况下的自动标引评价结果有较大的差异。无参照情况下的自动标引模型中，SVM自动标引模型的得分最高，而回归模型与基准模型的得分比较接近，CRF模型的得分最低。由于CRF模型在进行关键词抽取时，有较高的查准率，但召回率比较低，因此在无参照情况下，由于是将标引关键词这样的小集合与原文的概念向量（文本利用VSM简化表示）这样的大集合进行比较，因此得分出现偏低情况。

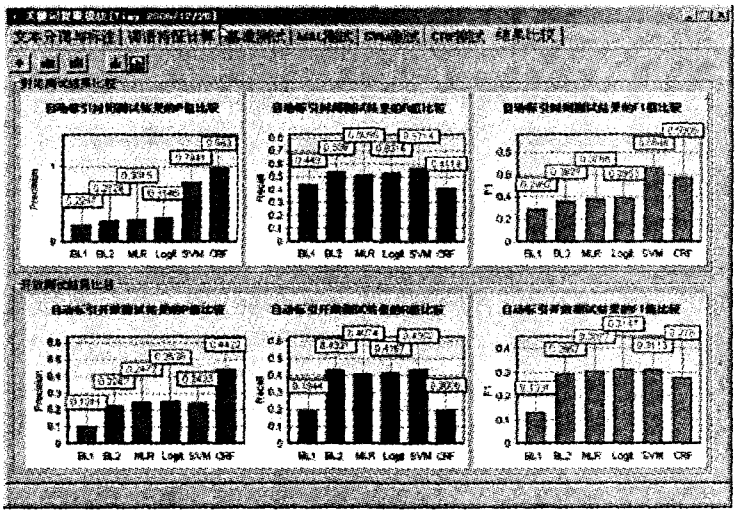

图4　自动标引传统评价结果

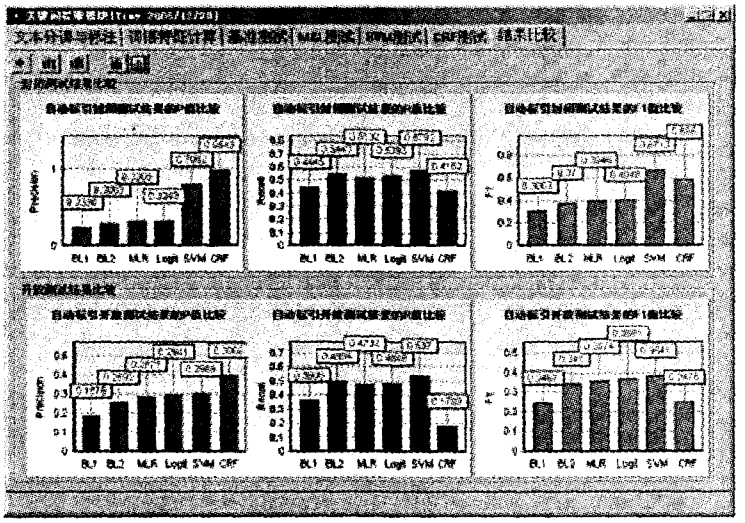

图5　自动标引通用评价模型评价结果

需要指出的，无参照情况下的评价结果比较依赖于原文“替代者”的质量。今后进一步的工作包括对原文的概念向量的生成与更加合理的相似度计算方法。

3.2　自动标引通用评价模型性能分析初探

本节对自动标引通用评价模型的性能分析，尝试从定性分析与定量分析相结合的角度进行初步探索。

3.2.1　定性分析

自动标引通用评价模型主要利用相似度原理考虑关键词之间的语义关联，克服了传统评价方法中仅考虑关键词精确匹配（造成

大量的赋词标引情况被忽略)这一缺点。

有参照情况下的评价模型属于内部评价中的标引结果比较法。本文提出利用主题词表、语义词典、本体等外部资源进行了关键词关联控制的方法,来对传统评价方法进行修正。

表3 相似度取值范围与标引结果得分的对应关系

相似度范围	0 - 0.19	0.20 - 0.39	0.40 - 0.59	0.60 - 0.79	0.80 - 1.00
标引结果得分	0	1	2	3	4

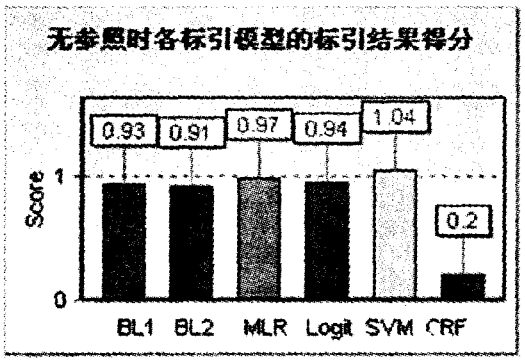

图6 无参照情况下的自动标引评价结果显示

无参照情况下的评价模型可以看成内部评价中的要点评价法。其主要思想是先提取论文的主要要点,得到原文的概念向量,将概念向量与标引结果进行相似度比较,根据比较结果来判断标引的质量。

值得注意的是,自动标引通用评价模型也存在一些需要解决的问题,如无参照评价中文本概念向量是否能很好地代表原文、有参照评价中参照关键词的强依赖性,这些都可能影响该模型评价结果的真实性。未来的工作还包括对文本表示模型进行探索,进行文本本体化的表示并用于标引结果的自动化评价,将是一个非常有趣的工作。

3.2.2 定量分析

本模型中,无论是无参照情况还是有参照的情况,都是以两个词语之间的相似度计算为基础。因此相似度算法是影响本模型性能的一个重要因素,本节对自动标引通用评价模型中所使用的相似度算法进行性能评价。

本文对相似度计算性能的评价是考察相似度计算的质量,即计算结果的真实性。本文将评价过程中关键词对的相似度计算结果单独提取出来,进行人工评估。评估方法为对这些相似度计算结果根据其真实性进行人工打分,根据打分结果,进行统计分析,总体评价相似度计算的可靠程度。其中打分依据如表4所示。

本文选择3.1小节所使用的600篇文献为研究对象,以6个标引模型进行自动标引,从获得结果中,随机选取其中500依靠相似度计算进行关联的关键词词对。表5为部分关键词词对与相似度计算结果样例。其中综合相似度是综合考虑语义相似度和词形相似度后的结果。本文设定语义相似度和词形相似度的权重的经验值分别为0.6与0.4。500对关键词的相似度评分结果统计如表6所示。

表4 标引评价打分表

关键词的似度计算结果是否符合实际	十分符合	基本符合	不太符合
人工评价得分	3	2	1

由表6可以看出,关键词相似度计算的结果中,符合实际情况的结果约为75%,说明了相似度计算方法具有一定的可靠性。在应用可以通过提高相似度阈值的方法来过滤相似度过低的关键词词对,但这样可能会遗漏一些可能在语义上比较相关的关键词词对。因此,提高关键词语义相似度的正确率是需要进一步深入研究的工作。

表5 关键词相似度计算样例

关键词1	关键词2	语义相似度	字面相似度	综合相似度
反倾销	倾销	1.0000	0.7444	0.8978
经济效应	贸易转移效应	0.7000	0.4132	0.5853
价值链	价值链分析	0.7000	0.6480	0.6792
国际分工	要素分工	0.6208	0.5800	0.6045
产业补贴	补贴政策	0.6389	0.5000	0.5833

表6 关键词的相似度评分结果统计

得分	3	2	1
数目	138	235	127
比率(%)	27.60	47.00	25.40

4 结束语

针对常规自动标引评价方法存在的评价结果不能完全反映真实标引结果，以及评价成本高的情况，本文提出一种通用的自动标引评价模型，并对模型的性能分析进行初步的探索。自动标引通用评价模型主要借助于外部的资源，根据相似度原理对传统测评方法中忽略关键词之间的语义关系这一个问题进行了修正。应用评价结果表明该通用评价模型具有一定的可靠性。

关于自动标引通用评价模型，今后的进一步的工作主要包括如下两个方面：

(1)融合更多种的资源提高关键词语义相似度计算的可靠程度

融合更多的外部资源，如主题词表、HowNet 或领域本体，进一步提高关键词语义相似度计算的可靠程度。另外利用大规模语料库进行基于关键词的语义相似度的计算、将多种方法进行融合并应用于自动标引评价中也是今后需要深入研究的课题。

(2)融合更多方法提高通用评价模型的可靠程度

在评价方法的融合方面，今后进一步的工作主要包括：有效利用标引结果的外部评价，即在使用标准的语料库上，通过自动标引结果在文本分类、文本聚类以及文本检索结果中的性能，快速地评价自动标引的质量。将内部评价与外部评价进行整合，提高自动标引通用评价模型的可靠性和适应性，使通用评价模型的评价结果接近真实情况下的评价结果。

参考文献

1 曾元显. 关键词自动提取技术与相关词反馈. 中国图书馆学会会报, 1997, 59: 59 - 64

2 李素建, 王厚峰, 俞士汶, 辛乘胜. 关键词自动标引的最大熵模型应用研究. 计算机学报, 2004, 27(9): 1192 - 1197

3 Chien L F. PAT - tree - based Keyword Extraction for Chinese Information Retrieval. In: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 1997), Philadelphia, PA, USA, 1997: 50 - 59

4 Turney P. D. Learning Algorithms for Keyphrase Extraction. Information Retrieval, 2000, 2: 303 - 336

5 Moens M F. Automatic Indexing and Abstracting of Document Texts. Boston/Dordrecht/London: Kluwer Academic Publishers, 2000: 78, 104

6 Zhang K, Xu H, Tang J, Li J Z. Keyword Extraction Using Support Vector Machine. In: Proceedings of the 6th International Conference on Advances in Web - Age Information Management Conference (WAIM2006), Hong Kong, China, 2006: 85 - 96

7 Turney P D. Mining the Web for Lexical Knowledge to Improve Keyphrase Extraction: Learning from Labeled and Unlabeled Data. Technical Report ERB - 1096. (NRC # 44947), National Research Council Canada, 2002: 1 - 34

8 贾同兴. 人工智能与情报检索. 北京：北京图书馆出版社, 1997: 19

9 Deerwester S, Dumais S T, Landauer T K, Furnas G W, Harshman R A. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 1990, 41(6): 391 - 407

10 Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing. Communications of ACM, 1975, 18(11): 613 - 620

11 侯汉清，章成志，郑红. Web 概念挖掘中标引源加权方案初探. 情报学报, 2005, 24(1): 87 - 92

12 Ristad E S, Yianilos P N. Learning String - edit Distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(5): 522 - 532

13 Medelyna O. Automatic Keyphrase Indexing with a Domain - Specific Thesaurus. Master Thesis, University of Freiburg, Germany, 2005: 29 - 32

14 Dagan I, Lee L, Pereira F. Similarity - based Models of Word Cooccurrence Probabilities. Machine Learning, 1999, 34(1 - 3): 43 - 69

15 章成志. 基于多层特征的字符串相似度计算模型. 情报学报, 2005, 24(6): 696 - 701

16 人大报刊复印资料. [2005 - 12 - 20]. http://www.confucius.cn.net

章成志 男，1977 年生，博士，讲师，目前于中国科技信息研究所从事博士后科研工作，主要研究方向为信息检索、数据挖掘与自然语言处理。

周冬敏 女，1984 年生，硕士研究生，研究方向为信息检索。

苏新宁 男，1955 年生，教授，博士生导师，研究方向为情报检索算法与中文信息处理技术。