一种基于最大匹配和向量空间模型的用户检索词规范化方法*

何伟[1,2]，常春[1]
（1.中国科学技术信息研究所，北京 100038；2.怀化学院，怀化 418008）

摘要：由自由词描述的用户检索词，可能会导致返回过多或过少的检索结果。有研究显示使用叙词表中的语词作为检索词，可提高网络检索系统的查准率和查全率。基于此，本文提出一种基于最大匹配和向量空间模型的用户检索词规范化方法，从词形和词义上进行规范化处理。首先使用最大匹配方法从词形上对用户检索词进行规范化；然后对用户检索词以及词形规范化后的语词构造词汇向量，计算它们间的语义相似性，从词义上实行规范化，获得最终的规范化语词。试验结果表明：本文提出的方法取得较好的效果，用户检索词返回的结果大部分都可通过规范化语词检索获得，当检索词为单个词语时，查准率超过90%。

关键词：最大匹配；向量空间模型；规范化；叙词表
中图分类号：G254
DOI：10.3772/j.issn.1673-2286.2016.7.006

1 引言

目前多数网络信息检索工具所采用的关键词检索，是一种不受控的自然语言检索机制，广泛存在“一义多词”和“一词多义”的现象，且用户由于知识背景、检索经验的不同，可能会选择不同的检索词进行检索，导致匹配失败而漏检。一些学术搜索引擎，如万方数据知识服务平台、百度学术搜索、中国知网等，虽然在搜索实现过程中使用一些关键词和受控词汇的标引，但大部分的标引词汇来源于用户经常使用的检索词。就目前的应用状况来看，用户难以用简单的检索词或词串准确地表达其信息需求，更不擅于利用多个检索词组配形成检索式。检索词的选择或检索式的构造不恰当，会导致返回过多或过少的检索结果。已有研究结果表明：使用叙词表中的受控词汇作为检索词，可为检索扩展形成概念检索奠定基础，进一步提高检索系统的查全率和查准率[1]。因此，探索用户检索词的规范化处理方法，对检索词进行受控词汇表述，可为进一步提高检索工具的效率提供便利条件。

关于检索词的规范化，国内外的相关研究成果并不多，多数都是从叙词表编制的角度关注文献关键词的规范化理论和方法，并作为检索系统的一部分来论述[2-3]。譬如，马张华等使用字面匹配相关控制方法将用户检索关键词与词集匹配，提供含有关键词成分的复合词，供用户选择[4]；李育嫦则从叙词表入手，阐述自然语言检索中，用户检索词词汇控制的必要性、基本措施和特点分析，但没有直接指出具体的检索词规范化方法[5]；宋斐雯等则在自然语言检索技术上对用户检索词进行概念语义上的控制，通过概念空间，对用户检索词进行语义处理，形成概念[6]；熊霞等介绍了一种基于扩展叙词表的检索关键词规范化方法[7]，该方法首先从网络信息和用户日志中抽取有实际含义的词补充到核心叙词表中，形成扩展型叙词表，然后将用户输入的检索词与扩展型词表进行完全匹配，通过扩展词表和核心词表间的映射，实现检索词的规范化，但没有进行实际验证；韩其琛等在其所构建的林业信息语义检索模

* 本研究得到中国博士后科学基金项目“基于叙词表语义关系的智能检索模型研究”（编号：2014M550791）资助。

型中，使用叙词表对用户检索词进行规范化处理，在规范化过程中，通过部分匹配技术返回与用户检索词部分匹配的结果，供用户选择新的检索词，若没有结果返回，则不进行规范化[8]。

在上所述方法中，多数是利用部分匹配技术，将用户检索词与词表中的词汇进行匹配，返回含有检索词成分的受控词汇供用户选择。此类方法只是在词形上进行简单匹配的规范化。基于此，本文系统提出并阐述一种基于最大匹配和向量空间模型的用户检索词规范化方法。该方法借助叙词表，利用最大匹配和向量空间模型从词形和词义上将用户检索词规范化为叙词表中的语词，并从检索词为单个词语和检索词为句子两个方面对该方法进行讨论分析，最后通过试验验证本文所提出的规范化方法。通过对用户检索词进行规范术语控制，可以对检索词进行扩展或优化，从而为进一步提高检索效率提供前提条件。

2 最大匹配和向量空间模型概述

2.1 最大匹配分词法

最大匹配是一种简单、便捷、易用的分词方法。它的具体策略为假设存在一个分词词典A，若词典A中最长词条有n个字符，则取待分词的字符串B的前n个或后n个字符作为候选分词字段，在词典A查找匹配的词汇。若在词典A中存在这样的一个n字词，即匹配成功；否则，将候选字段的最后（前）一个字符去掉，重新在词典A中进行匹配，如此反复分割匹配，直至匹配字段的长度为0，即匹配不成功。否则，匹配成功，按照字典A中匹配的词对语词B进行分词[9]。

借用最大匹配分词策略的原理，可以将叙词表看作分词词典，用户检索词看作待切分的字符串，可利用最大匹配分词方法将检索词在叙词表中进行最大匹配，获得在词形上最接近用户检索词的叙词表语词，从而对用户检索词在词形上进行规范化。

2.2 向量空间模型

向量空间模型（Vector Space Model）是Salton等于1975年提出的一种高效、简洁的文本表示模型，通过提取文本中的特征词，构造文本向量来描述文档信息[10]，利用向量计算检索词与待检索文本的语义相似性。若两个向量间的相似性值越大，则说明向量所表示的语词间的语义越相似，其语义关联程度越大；若相似性值为0，则说明两语词间的语义不相关。

因此，利用向量空间模型为通过词形规范化后获得的语词和用户检索词分别构造词汇的向量表示，通过向量计算叙词表语词与用户检索词间的语义相似性，得到在语义上与用户检索词相近的叙词表语词，从语义上对用户检索词进行规范化。

3 用户检索词规范化方法

本文提出的用户检索词规范化方法分为两步。第一步，利用最大匹配方法将用户检索词与叙词表中的语词进行最大匹配，获得与用户检索词在词形上最接近的叙词表语词候选集，对用户检索词进行词形上的规范化；第二步，为用户检索词及每个候选词构造词汇的向量表示，并计算它们间的相似性，完成用户检索词语义上的规范化。

3.1 规范化实施过程

本文提出的基于最大匹配和向量空间模型的用户检索词规范化流程见图1。

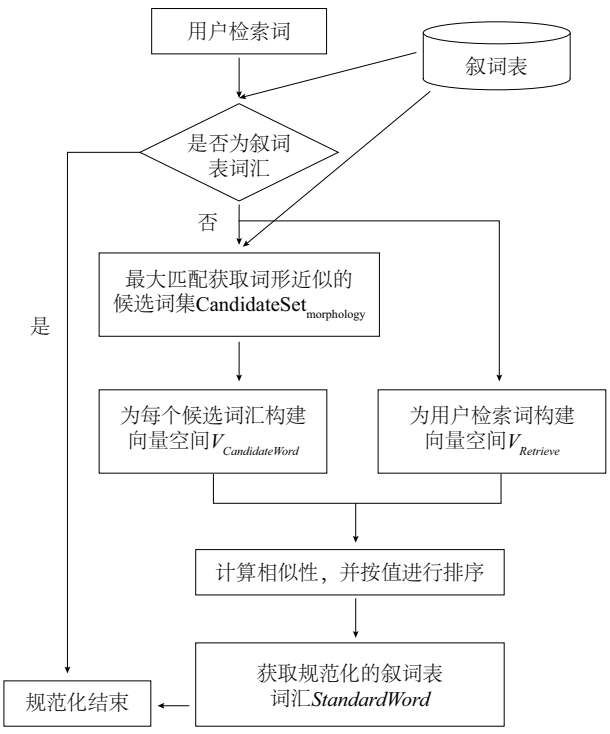

图 1 用户检索词规范化流程

具体流程：

（1）对于用户提交的检索词，与叙词表中语词进行匹配，判断该检索词是否为叙词表中的语词，若是，则规范化结束，并输出规范化后的检索词；否则，利用最大匹配方法在叙词表中查找在词形上与用户检索词最接近的叙词表语词，形成词形规范化候选词集，记为$CandidateSet_{morphology}$；

（2）对候选词集中的每一候选词汇*CandidateWord*和用户检索词*Re_Word*，使用中国知网，获取其相关检索词，分别构建候选词词汇向量表示$V_{CandidateWord}$和用户检索词词汇向量表示V_{Re_Word}；

（3）利用公式（1）计算两个词汇向量的相似性，并对相似性值进行排序，公式（1）表明若两个向量含有的共同词汇越多，则这两个向量的重合程度越高，越相似；

$$Sim(CandidateWord, Re\_word) = \frac{|V_{CandidateWord} \cap V_{Re\_Word}|}{Max(|V_{CandidateWord}|, |V_{Re\_Word}|)} \quad (1)$$

公式（1）中$V_{CandidateWord}$，V_{Re_Word}分别为候选词和用户检索词的词汇向量表示，$|V_{CandidateWord}|$为向量$V_{CandidateWord}$的词个数。

（4）选择相似性值最大的候选词为用户检索词的规范化语词*StandardWord*，实现对用户检索词的语义规范化处理，获得规范化语词，完成用户检索词的规范化过程。

3.2 用户检索词的分析处理

（1）用户检索词为单个词语。当用户提交的检索词为单个词语时，如“职业教育”，则对其的规范化使用本文中所述的过程进行处理。

（2）用户检索词为句子。当用户提交的检索词为句子时，如“论职业技术教育的本质属性”，则需在规范化前，对其进行分词、停用词过滤等操作，获取该句子中的实词，并根据具体情况对这些实词进行逻辑组配（并或交操作）形成用户检索词的逻辑表达式。譬如，将上例中的检索句转换为“职业技术教育∩本质∩属性”的逻辑表达；然后对每个实词使用本文所述的规范化过程进行处理，获得每个实词的规范化语词；最后按照原来的逻辑操作（并或交）将这些规范化语词重新组合，完成检索句的规范化过程。

3.3 用户检索词规范化算法

根据前面的分析和讨论，本文设计了一个基于最大匹配和向量空间模型的用户检索词规范化算法。

输入：用户检索词*Re_Word*，叙词表T；

输出：规范化后的叙词表中语词*StandardWord*；

Begin

Step1：提交用户检索词*Re_Word*，在叙词表T中判断该检索词是否存在，若存在，则转到Step6；否则转到Step2；

Step2：利用最大匹配方法在叙词表中为用户检索词*Re_Word*获取在词形上最接近该词的候选词集$CandidateSet_{morphology}$，

$CandidateSet_{morphology}$={CandidateWord_1, CandidateWord_2,…, CandidateWord_n}；

Step3：使用中国知网获取用户检索词*Re_Word*以及候选词集$CandidateSet_{morphology}$中每一词汇*CandidateWord_i*的相关搜索词，分别构建用户检索词词汇向量V_{Re_Word}和候选词词汇向量$V_{CandidateWord_i}$；

Step4：使用公式（1）计算用户检索词*Re_Word*以及候选词集$CandidateSet_{morphology}$中的每一词汇*CandidateWord_i*的相似性，并进行排序；

Step5：获得相似性值最大的候选词*CandidateWord*作为用户检索词*Re_Word*的规范化语词*StandardWord*，并输出；

Step6：规范化结束。

End

4 试验与结果分析

为验证本文提出的检索词规范化方法的有效性，本文使用中国知网对其进行验证[11]。叙词表采用《综合电子政务主题词表试用本范畴表》（以下简称“政务主题词表”），该词表是由中国科学技术信息研究所于2005年1月编制完成的我国第一部按国家标准编制的综合性电子政务主题词表，共收录主题词20 252条，其中正式主题词17 421条，非正式主题词2 831条；范畴索引划分为21个大类，132个二级类[12]。“政务主题词表”由电子政务名词术语经语义相关、概念等级相关和知识领域相关处理后选编而成的规范化动态性词表，其中术语间语义关系主要通过符号“Y”代表“用”，“D”表示“代”，“S”代表“属”，“F”表示“分”，“C”代表“参”[12]。

假设用户提交的检索词为“职业教育”，在“政务主题词表”中进行查找，若发现该检索词不存在，则需要规范化。利用最大匹配法在“政务主题词表”中分别进行正向和逆向匹配，得到$CandidateSet_{morphology}$={职业技术教育}；使用中国知网分别获取“职业教育”和“职业技术教育”的相关搜索词，构建它们的词汇向量，分别为$V_{职业教育}$={职业教育，职业准备教育，德国职业教育，中国职业技术教育，职业技能教育，职业学校，职业技术教育，职业教育发展，中等职业学校，高职教育，职业教育课程，职业教育体系，职业教育研究，职业教育政策，职业教育集团}和$V_{职业技术教育}$={职业技术教育，职业技能教育，中国职业技术教育，职业技术教育改革，职业教育，职业技术学校，职业学校，高职教育，职业技术教育学，职业教育发展，职业教育研究，职业技术教育培训，职业技术教育管理}；计算“职业教育”和“职业技术教育”的相似性：

$$Sim(职业技术教育,职业教育)=\frac{\left|V_{职业技术教育}\cap V_{职业教育}\right|}{Max\left(\left|V_{职业技术教育}\right|,\left|V_{职业教育}\right|\right)}=\frac{8}{15}=0.534；$$

最后使用叙词表中的语词“职业技术教育”对用户检索词“职业教育”进行规范化，并作为新的检索用词进行检索。

在中国知网中，使用原用户检索词“职业教育”，不限学科领域进行主题检索，共得到421 895条数据记录（截至2016年5月31日），其中前20条记录见图2。

图2 “职业教育”主题检索前20条记录截图

在相同条件下，以“职业技术教育”为检索词，共获得398 743条数据记录，采用原用户检索词“职业教育”在该检索结果中二次检索，获得387 488条记录，为原检索结果的91.84%。将“职业教育”主题检索结果的前20条记录放入“职业技术教育”的主题检索结果中的“在结果中检索”或“职业技术教育”的全文检索结果中的“在结果中检索”，结果见图3。

图3 “职业教育”在“职业技术教育”的主题或全文检索结果的“在结果中检索”结果

比较图2和图3可见，利用原检索词“职业教育”进行主题检索获得的前20篇文章中，1篇可在“主题”——“职业技术教育”的检索结果中“在结果中检索”二次主题检索得到，16篇可在“全文”——“职业技术教育”检索结果中“在结果中检索”二次主题检索获得，1篇可在“全文”——“职业技术教育”检索结果中“在结果中检索”二次全文检索获得，其中2篇检索不到。可见，利用规范化的“职业技术教育”检索词替代原用户检索词“职业教育”进行检索可得到大部分检索结果，虽然可能会缺失小部分结果，但也产生了一些用户可能感兴趣的新的检索结果。

为进一步验证本文提出的用户检索词规范化方法，随机选取20个用户检索词（包含15个单个词语和5个句子），分属教育、计算机、经济、哲学、军事五个不同的领域；以中国知网作为检索系统，使用《综合电子政务主题词表试用本范畴表》为叙词表，每个用户检索词取返回的前50条记录进行分析，主要分析这50条记录中有多少条存在于对应规范化检索词的返回结果中。采用信息检索中的查准率作为评价指标对规范化前后的检索词返回结果进行评价分析，试验结果见表1。查准率的计算公式如（2）所示：

$$Precision=\frac{\left|ResultSet_{Re\_Word}\cap ResultSet_{StandardWord}\right|}{\left|ResultsSet_{Re\_Word}\right|}\times 100\%\quad(2)$$

其中$|ResultSet_{Re_Word}|$代表用户检索词在中国知网中返回的结果数，如果多于50，则取值为50，否则取实际的返回数量，$ResultSet_{StandardWord}$表示规范化后的检索词在中国知网中返回的结果集。

表 1 检索词规范化方法试验结果

用户检索词		规范化后的检索词	用户检索词返回的检索结果数量（前50）	存在于规范化后检索词返回的检索结果中的数量	查准率
检索词为单个词语	职业教育	职业技术教育	50	46	92%
	教育理念	教育理论	50	44	88%
	教育变革	教育改革	50	47	94%
	技术革新	技术创新	50	48	96%
	信息资源检索	信息检索	50	50	100%
	语义搜索	语义检索	50	50	100%
	信管	信息管理	50	50	100%
	边际成本	生产成本	50	36	72%
	金融危机	经济危机	50	45	90%
	消费支出	支出	50	47	94%
	马克思主义中国化	马克思主义哲学	50	43	86%
	自然辩证法	辩证法	50	48	96%
	民兵预备役	预备役	50	50	100%
	人民武装斗争	武装斗争	50	47	94%
	反恐行动	反恐怖行动	50	50	100%
检索词为句子	高校教师应具备的能力	高等学校教师、能力	50	36	72%
	基于叙词表的语义搜索	叙词表、语义检索	15	11	73.34%
	经济管理与经济发展的关系	经济管理、经济发展、关系	50	40	80%
	马克思主义中国化进程	马克思主义哲学、进展	50	35	70%
	反恐行动的若干问题	反恐怖行动、问题	8	6	75%

“存在于规范化后检索词返回的检索结果中的数量”的含义为用户检索词返回的前 50 条记录在规范化检索词后，返回结果中的数量。从表 1 可见，在规范化后的检索词返回结果中并没缺失太多的原检索结果，且增加了一些潜在的用户感兴趣的结果。对于用户检索词为单个词语的情况，得到的查准率大部分超过90%，但其中有一个用户检索词“边际成本”，规范化后获得的查准率值为72%。分析其原因，在对该检索词的最大匹配过程中获得的词形规范化候选词集为{成本，储蓄成本，科研成本，流通成本，生产成本，运输成本}，通过构造词汇向量表示并计算其相似性后获得的规范化语词为“生产成本”，这两个词的语义相似性仅为0.102，该词在含义上与“边际成本”有一定的出入，这可能是由于“政务主题词表”含有经济领域的词汇相对较少，无法获得语义非常相似的规范语词造成的。而在用户检索词同义语词的规范化中，查准率达到100%，且规范化后的语词获得的结果相对较多，这说明人们在论文写作时更倾向使用规范的语词，如“信息资源检索”和“信息检索”“信管”和“信息管理”“反恐行动”和“反恐怖行动”等。在一些用户检索词的上位概念规范化转换中，查准率也获得较高的分值，如“消费支出”和“支出”“自然辩证法”和“辩证法”“民兵预备役”和“预备役”等，这也间接地说明较宽概念的检索结果包括较窄概念。当用户检索词为句子时，获得的结果都在70%—80%，并不是很理想。这可能是由于分词后对句子的各个词汇进行规范化再重新组合时无法整合成句子，丢失了一部分语义信息，造成与用户表达有较大的差别。

从以上分析可知，本文提出的用户检索词规范化算法在很大程度上是有效的、合理的，特别是在检索词为单个语词时获得了较好的结果，表现出较强的竞争力。

5 结论

本文提出一种基于最大匹配和向量空间模型的用户检索词规范化方法，从词形和词义上对用户检索词

进行规范化转换，在尽可能少的语义损失下，从叙词表中获得合理的语词进行检索，为检索系统的扩展检索提供前提条件。试验结果表明：本文提出的方法是一种有效的检索词规范化方法，但也存在一定的不足。尤其检索词是句子时，所获得的效果没有单个词语好，这是因为对句子中的词汇进行规范化后，再进行重组导致部分信息损失。下一步工作将对这种情况进行深入地研究，考虑句子重组的规则策略。

参考文献

[1] 薛春香,侯汉清.叙词表词汇控制机制变革的探讨[J].图书馆杂志,2013(11):38-44.

[2] VERHODUBS O. Towards the ontology web search engine[J].Addiction, 2015,110(S2):54-58.

[3] GÖDERT W. An ontology-based model for indexing and retrieval[J]. Journal of the Association for Information and Technology, 2016,67(3):594-609.

[4] 马张华,李玲.文本检索中的词汇控制研究[J].图书和情报工作,2004,48(21):84-87.

[5] 李育嫦.自然语言检索中的词汇控制研究[J].图书馆学研究,2006(4):75-78.

[6] 宋斐雯,王洋.自然语言检索中的概念语义控制[J].计算机时代,2011(2):4-7.

[7] 熊霞,胡秀梅,杨江丽.网络信息检索中传统叙词表的分析与改进[J].四川图书馆学报,2012(3):24-26.

[8] 韩其琛,李冬梅.基于叙词表的林业信息语义检索模型[J].计算机科学与探索,2016,10(1):122-129.

[9] 闻玉彪,贾时银,邓世昆,等.一种改进的最大匹配中文分词算法[J].计算机技术与发展, 2011,21(10):92-94,98.

[10] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J].Communications of the ACM, 1975,18(11):613-620.

[11] 中国知网[EB/OL][2016-05-31].http://epub.cnki.net/kns/brief/result.aspx?dbprefix=scdb &action=scdbsearch&db_opt=SCDB.

[12]《电子政务主题词表》编制与应用系统课题组.综合电子政务主题词表试用本范畴表[M].北京:科学技术文献出版社,2005.

作者简介

何伟，男，1978年生，博士，研究方向：本体工程，语义计算。

常春，男，1966年生，博士，研究馆员，研究方向：信息组织。

An Approach for Normalizing Retrieval Word Based on Maximum Matching and Vector Space Model

HE Wei[1,2], CHANG Chun[1]

（1.Institute of Scientific and Technical Information of China, Beijing 100038, China;
2. Huaihua University, Huaihua 418008, China）

Abstract: It can conduct much more or a little result using free terms as retrieval word. Existing research results show that it can improve the recall and precision of a retrieval system using normalized terms from controlled vocabularies. In this paper, we propose a new approach to normalize retrieval words base on maximum matching algorithm and vector space model, which deal with the retrieval words in the two aspects of morphology and semantics. This method first exploits maximum matching to normalize the retrieval words from morphology and obtain candidate words, then respectively construct the vector of the candidate word and the retrieval word to compute semantic similarity, and finally selected the most similar candidate word as the normalized word of the retrieval word. The experimental results showed that the proposed method obtained a promising result, with the precision of more than 90% on the condition that retrieval word is a single word.

Keywords: Maximum Matching; VSM; Normalization; Thesaurus

（收稿日期：2016-07-05）