基于词汇链的文本过滤模型*

尤文建，李绍滋，李堂秋
(厦门大学 计算机系，福建 厦门 361005)

摘 要：在介绍文本过滤的背景及传统基于关键词的向量空间方法不足之处的同时，引入了词汇链的概念，提出了基于词汇链表示文本的文本过滤模型，该模型首先对文本进行分析，把文本表示成词汇链的形式，在形成用户初始模板之后，以此模板来进行文本过滤。在用户反馈的基础上，自适应地修改该模板，以适应用户变化的需求及改善系统过滤性能，实验表明，这样的确可以提高系统精度。

关键词：词汇链；向量空间；用户模板；文本表示；WordNet
中图法分类号：TP391 **文献标识码**：A **文章编号**：1001-3695(2003)09-0032-04

A Text Filtering Module Based on Lexical Chain

YOU Wen-jian, LI Shao-zi, LI Tang-qiu
(*Dept. of Computer Science, Xiamen University, Xiamen Fujian* 361005, *China*)

Abstract: In this paper we first give some information about the text filtering and the defects in VSM(Vector Support Machine), then we introduce the concept of lexical chain, give a model that build a profile based on lexical chain. We first analyse the text, then express the text with lexical chain. We use this lexical-chained profile to sift the information which may be of the user's interest. A filtering system should be able to adapt to user's interest changes, so we automatically modify the user model to recognize the changes. Experimental results show that the methods can improve the text filtering performance.

Key words: Lexical Chain; Vector Space; User Model; Text Representation; WordNet

1 引言

Internet 的迅速发展，已使它成为全球最大的分布式信息库。用户在享受它方便和快捷的同时，也为它所包含的庞大芜杂的信息所淹没，往往为了找到自己需要的信息花费大量的时间和精力。如何能够更有效、更准确地找到自己感兴趣的信息，过滤出与自己的查询需求有关的信息已成为基于 Internet 网络信息的当务之急。由于目前绝大多数信息均表现为文本方式，所以有关文本处理的各种技术得到极大的促进和发展。如文本摘要、文本检索和文本浏览等，这些为计算语言学的发展注入了新的活力，也是计算语言学能够获得实际应用并且取得良好效果的领域，这引起了国际上学术界、企业界的研究人员极大的兴趣。在著名的文本检索会议(Text Retrieval Conference, TREC)中，文本过滤已经成为其中一个重要的议题。TREC 在信息过滤的理论和技术研究以及系统测试评价方面，对信息过滤的形成和发展提供了强有力的支持。

早在 1987 年的时候，Malone 及同事基于他们的“Information Lens”系统，引入了三种信息过滤模型：认知的(Cognitive)，经济的(Economic)，社会的(Social)。后来 Denning 对认知的模型更明确地定义为：基于内容的过滤模型(Content-based Filtering)[1]。由于该方法假定用户的操作都是独立进行的，因此基于内容的过滤的文本表示能利用的只是来源于文本内容的信息。基于内容的过滤模型已经成为主要的方法，很多人在这方面做了许多工作[2,3]，但是他们的主要方法都只是从文本中抽取一些关键词，他们只孤立地考虑各个关键词，而忽略关键词之间的关联。这可能使表示出来的文本与实际文本之间差别比较大，脱离实际文本的主题，从而过滤出许多与主题无关的文本。

本文以语义词典 WordNet 为知识来源，引入词汇链(Lexical Chain)的概念。在分析了关键词之间以及主题与词之间的关系后，给出了词汇链的构造方法、评价方法，并用它来表示文本的内容，使其更加真实地反映了文本的实际主题内容，并用它来进行文本过滤，很好地解决了以前文本过滤方法的一些不足之处。

2 基于内容的文本过滤模型系统结构

基于内容的文本过滤工作基本上可以分为以下几部分：①对用户的要求进行认知，即把用户的要求在计算机内表达；②对输入文本进行表达，以便能够与用户

收稿日期：2003-01-12
基金项目：国家“863”计划项目(2001AA114110)；福建省科技计划重点项目(2001H023)

的要求进行比较,把符合要求的文本提供给用户;③对输入文本进行判断,即把用户的要求与输入文本进行比较,然后把输入文本划分为相关文本或无关文本。

文本过滤的主要流程是根据用户的信息需求,建立用户需求模型;然后在相应的文本流中过滤出符合用户需求的文本提交给用户,由用户对文本作出是否符合其需求的评判,再根据评判的结果自适应地修改用户模板,更好地符合用户的需求。

由于向量空间模型具有表示简洁和计算简便的特点,因此,在文本检索、文本过滤和文本摘要等方面获得广泛应用,取得了一定的效果。一般是从文本中抽取出关键词,根据该词在文本的重要性,给每个词赋予一定的权重,把用户模板和未知文本均表示成向量空间中的向量,利用它们的夹角的余弦来进行相似度的度量。但是在提取关键字之后,以前的方法只是根据词语在文本出现的频率赋予相应的权值。这里我们将分析以前方法的一些不足,并给出了改进的方法。

2.1　文本表示方法的改进

传统的信息过滤技术主要是采用关键词查找和统计技术来过滤相关的文本。只是分别计算各个词在文本中出现的频率,而忽略词之间的语义关系,忽略主题与词之间的关系。对于这种把关键词孤立起来的技术来说具有很大的局限性。这里我们分析传统关键词技术不足的三种主要因素并给出相应的解决方法。

(1) 无用词的影响

就是各个类别中均可以出现的特征,它不代表类别的特点。这些词有些是属于停用词处理,还有就是那些信息量不大的词,对于这些词我们将删除。

(2) 词间关系的影响

可分两种情况说明:①同义词的影响(如计算机与电脑);②具有某种语义关联词的影响,如医疗类中,“医生”、“护士”、“医院”、“病床”、“手术室”、“诊断”、“药方”、“感染”、“病情”、“抗体”等词是存在某种关联的。其中一个特征的存在在某种程度上具有替代其它词的作用,各个特征单独出现的频率可能比较小,而且也许会被一些无关的、出现频率大的词所覆盖。对于这种情况,我们可以考虑词之间的语义关联,如果这些词共同表达的是一个主题的话,它们出现在词典中的语义距离是比较近的,在文本分析过程中就可以自动地放在一起,而在计算文本相似度的时候就可以把它们综合起来考虑。

比如从文本中抽出这样一些词信息如下:{{information:3,technique:1,Bayesian-technique:1,datum:2,model:1,rea:1}{computer:4}}

其中每个词后面的数字表示在文本中出现的次数。

如果只是分别考虑各个词的词频的话,则 Computer 最高,但是我们可以知道前面几个词之间有很强的语义关联,它们可以相互补充,从而提高该部分各个词的重要性。

(3) 词间地位不平等性的影响

关键词对主题支持的作用大小尽管可以用出现次数加权值大小来体现,但我们觉得还不够。人在看文章时并不一定要阅读全文,常常在读了标题或第一段后就可以较准确地确定主题。这说明,存在一些特征词对某一主题具有较强的支持作用(决策特征),它们的存在可以在很大程度上决定文本的主题。而在向量空间模型中,这种决策作用将可能被众多非决策特征的影响所淹没掉。对于这种情况,我们引入了特征区域概念。

文本特征区域是文本能够体现文本主题的区域,包括标题、摘要、关键字、参考文献。但并不是所有的文本主题都有摘要、关键字和参考文献,因此这些结构单元作为可选的单元。国内有人抽样统计,国内中文期刊自然科学论文的标题与文本的基本符合率为 98 %,新闻文本的标题与主题的基本符合率为 91 %[6]。任何文章几乎都有标题,因此标题是主要的文本特征之一。

线索词是那些总结性或是概括性的标志性词语,比如“总之”、“总而言之”、“综上所述”等。我们将加强特征区域所包含词的重要性,同样,对于线索词之后的关键词,我们也将增加权重,从而突出该词的重要性。

基于以上的考虑,我们想以 WordNet 为知识来源,在文本分析的时候,自动构造文本的词汇链。分析这些词汇链,从而得到真正文本内容的表达。

2.2　词汇链及其应用

词汇链(Lexical Chain)是指一个主题之下的一系列相邻的词之间共同组成的词系列。这些词在同样的词法环境中共现,因为它们描述的事情是一样的[7]。

比如{postoffice,service,stamps,pay,leave}这些词在特定的服务环境中会共现,因为它们都是描述服务的场景。这些词之间存在距离,而且它们共现是在一定的范围内的,但它们并不仅限在句子之内,词汇链描述的是文本中语义单元。在文献[7]中,Morris 和 Hirst 最初引进 Lexical Chain 概念的目的是文本分割,就是得到文本的结构。当时的基本想法是:因为词汇链是一系列相关词组成的,这些词表达的是同一件事情或意思,找到这些链就得到了文本的结构。不同的链条就构成了对该文本的结构分割。后来很多人根据这一基本想法,在很多方面得到了应用,比如在文本检索[8]、信息抽取[10]、检查文本的用词不当[9]等。

基于以上的认识,我们将采用词汇链来表示用户模板和未知文本。首先把用户提交的需求进行分析,然后构造词汇链来表示文本,从而构建出用户模板来表达用户的需求。这个需求在过滤的过程中采用反馈进行自动学习,使其更精确地满足用户要求。我们对未知文本,也采用同样方法构建出词汇链来表示文本。

美国普林斯顿大学认知科学实验室的米勒(G. A. Miller)和贝克威斯(R. Beckwith)等人,于 1985 年开始致力于建造词汇关系网络的工作,建成了 WordNet。WordNet 是一个在线词汇参照系统(在网上可机读的英语词库),是一个基于心理语言学原则的机器词典。WordNet 目前包括了大约 95 600 个词条,其中包括单纯词 51 500 个,复合词 44 100 个。WordNet 用同义词词集(Synsets)来表示词义,这 95 600 个词构成了由 70 100 个词义组织而

成的同义词集[4]。

这样,我们在对文本分析之后,可以利用语义词典 WordNet 提供的知识,根据 WordNet 中的语义关系构建文本的词汇链表示,这样就可以更准确地表示文本的内容,改善过滤系统的性能。系统结构如图 1 所示。

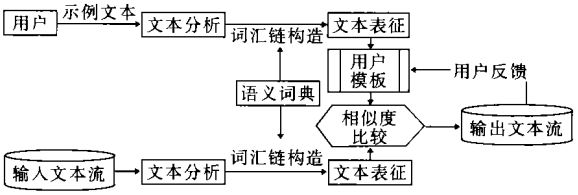

图 1 基于词汇链的过滤模型

3 过滤模型的设计和实现

3.1 文本表示方法

(1) 文本分析

并非文件当中所有的词都用于构造文本的词汇链,只有那些最能代表文件所要表达的意思的词,也就是关键词汇可被用来构造。文本表示方法可归纳如下:①文本预处理。对词根进行抽取,短语的识别等。②词性标注。对文本中的单词进行词性标注。③关键词抽取。在文本中去除所有属于下列的单词:冠词(如 a ,the ,an)、介词或连接主句和从句的副词(如 in ,to ,of)、情态动词(如 would ,must)和连接词(如 and)等,我们用 W(s ,w ,c)表示。其中 w 表示这个词,s 表示这个词在本文本中的序号,c 表示该词的词性。比如,(12 ,think ,verb)表示在 W 中第 12 个词是 think ,这是一个动词。我们也可以给各种词性的词赋予不同的权值来表示它们不同的重要性,一般而言,名词要赋予最大的权值。对于那些在标题、首段、末段、段首、段尾出现的词语也可以增加其权重。我们也可以设一个阈值,把那些出现频率低于该频率的词去除。④词汇链表示文本。在经过了关键词提取之后,得到文本的词的系列,再经过词汇链的自动构建方法,得到文本的词汇链表示。词汇链中的各个关键词的初始权重为该词出现在文本中的频率、词性加权系数、关键词之间的相关权重及关键词出现在特征区域的权值共同组成。

(2) 构建词汇链的方法

构建的时候是以词典中的词汇关系来自动完成的,因此主要考虑以下的几种词汇关联:

● 超强关联(Extra-strong)指两个词词义上的重复,两个词的间距不考虑。

● 强关联(Strong)可以分为三种情况(两个词间距设定为一个窗口,一般取七个句子):①两个词共同出现在同一个同义词集中,比如 human 和 person 它们都出现在同一个集合中{person ,individual ,someone ,man ,mortal ,human ,soul}。②两个词的同义词集中的某两个集合存在着某种语义关系(同义、反义、相似等),比如 precursor 的一个同义词集{predecessor ,precursor ,antecedent }与 successor 的一个同义词集{successor}存在着反义关系。我们称这种关系为水平连接关系。③如果有一个是短语或是复合词,而另一个词的某个同义词集中的词就包含在短语或是复合词中,在这种关系中我们并不考虑这种包含是一种什么样的关系。比如 private _ school 的同义词集{private _ school},而 school 的一个同义词集{school}就被另一个词的同义词集所包含,而且这种包含是一种上下位的关系。

● 中等关联(Medium-strong)指当存在着一种路径连接着两个词(两个词间距设定为一个窗口,一般取三个句子)。比如 apple 的同义词集中{apple}的上位词{apple} @{fruit} @{product ,green _ goods}～{vegetable ,veggie}～{carrot}最后得到 carrot ,说明有一条路径连接这两个词。而且我们限制这种路径的长度是在 2～5 之间,如果不在这个范围之内就不是这种关系了。

关键词之间的相关权重的计算:

weigth = A-path _ length-B * number _ of _ changes _ of _ direction (A 和 B 都是常数)

(3) 词汇链权重的计算

我们在对文本中各条词汇链重要性进行计算的时候,将考虑这几个因素:①词汇链的长度,就是链条包含的词的数目。②构成词汇链的各个词初始权重。③词汇链覆盖文本的范围,能够完整构造该词汇链的文本的范围,覆盖的范围越大,则包含主题的内容就越多。④词汇链的分布密度,就是构造该链的词分布密度,节点分布越集中,整体的重要性就提高。⑤构造出的这条链的拓扑结构,考虑词之间关联程度,加强核心节点的重要性。

至此完成了文本表示成词汇链形式的构建,并对此构成文本的词汇链进行了评价,赋给相应权值。每个文本表示成 $T = \{T_1 , T_2 , \cdots , T_n\}$。其中 T_i 表示各个词汇链的权值,$T_i = \{t_{i1} , t_{i2} , \cdots , t_{im}\}$,而 t_{ij} 表示构成词汇链的关键词的权值信息。文本中每个词汇链权值越大,表达文本主题越强;反之,权值越小,离主题就越远。我们设定一个阈值,从中取出最强的几个链条来共同表示文本。

3.2 用户模板表示

与把文本表示成为词汇链一样,用户模板的初始表示也是采用词汇链形式来表示。由用户提供训练材料,材料可以包括用户所感兴趣的一些关键词、文章摘要或者文章,但是这些材料要属于同一个类别之内的材料,表示用户某一方面需求。如果用户提供的是摘要或文章,则需要像前面介绍的那样,把它们表示为文本的词汇链;如果用户提供的是关键词,则根据这些词直接把它表示成词汇链,所有这些词汇链经过自适应学习算法形成了用户初始模板 $V\{V_1 , V_2 , \cdots , V_n\}$。

当然我们可以为用户所给的文本赋予不同的权值来表示其不同的重要性。我们按如下公式给每一文本赋予不同的权值:$\text{weight}(\text{text}) = \log(N)$

其中 N 表示文本长度,我们还设置了一个阈值,所有小于该阈值的分量均赋值为 0,以此实现降噪处理。

3.3 文本过滤

至此为此,文本与用户的需求已表示成词汇链的形

式,文本与用户需求的相关度可以通过以下计算的余弦值来衡量:$\cos(a) = \frac{\sum_{ij} V_i \cdot T_j}{\sqrt{\sum_i V_i^2 + \sum_i T_i^2}}$

在需过滤的所有文本当中,我们可以根据这个值来进行相关度排序反馈给用户,也可以设定一个阈值 k。当某文本与用户需求的相关度大于 k 时则认为该文本符合用户需求,把文本按相关度大小的顺序返回给用户,把低于该值的所有文本去除或存在某处以备用户在有空时处理。我们可以把用户的反馈考虑进去,若用户认为几乎所有我们过滤出的文件都是他所感兴趣的,则我们可调低 k 值,反过来,若有很多文本不符合用户的兴趣,则我们调高 k 值。

3.4　自适应学习算法

随着文本过滤的进行,用户的需求可能会发生变化,系统如何适应这种变化,这就需要有一个自适应学习的过程。利用用户的反馈信息,及时地对用户模板进行自适应性的修改,以适应用户的变化,提高系统过滤性能。因此我们的用户模板的建立是采用自适应的聚类的方法。开始的时候用户提交少量的示例文本,随着过滤的进行,用户从输出中进行判断,把过滤出来的相关文本反馈给过滤系统,过滤系统根据这些反馈的文本自动地修改用户模板。这种方法有以下几方面优点:

(1) 开始的时候用户仅需要提交少量的示例文本,解决用户获取示例文本的困难。

(2) 自适应性,这种方法建立起来的用户模板,可以随着用户提交示例文本的改变,而自适应地修改。

该算法的基本思路是:对于新反馈的文本,将按照一定的权值加入到原来用户模板,而且我们规定它只是对原来模板的微调。具体的算法步骤如下:①把文本表示成词汇链的形式;②对于第一个文本的词汇链 $T = \{T_1, T_2, \cdots, T_n\}$;③对于每个新提交的示例文本 $C = \{C_1, C_2, \cdots, C_n\}$;$C = (1-\lambda) \cdot T + \lambda \cdot C$;④如果系统适应慢,提高 λ 值;如果系统适应过快,减小 λ 值。其中自适应参数 λ 根据我们的实验证明,开始时一般取 $\lambda = 0.2$。

4　过滤模型的实验结果及实验分析

本文采用 TREC-9 上的医学语料库 OHSUMED,这是著名的国家医学图书馆(National Library of Medicine)的 MEDLINE 医学文献库的一个子集,由 1998 年～1991 年的医学文献组成,共含有文本 348 566 篇,来自 270 种医学期刊,总容量为 400MB。其中 1987 年的文摘将作为训练语料,而 1988 年～1991 年的文摘将作为测试语料[5]。

实验结果表明用传统向量空间的方法平均精度为 38 %,而采用基于词汇链的方法平均精度为 47 %,后者比前者高 9 %,相比较过滤的精度提高了,改进效果非常好。这是因为我们的表示方法更加准确地表达了文本主题内容,从而提高了过滤的精确度。

5　结束语

从网络信息服务需求出发,我们认为信息过滤将会越来越为人们所重视,然而传统的文本表示方法中忽略了许多因素,造成文本表示中信息的丢失,影响文本过滤的性能。如何更好地表示出文本的内容,是我们需要考虑的重要课题。本文提出基于词汇链的文本表示方法,解决了以前孤立地对待关键词表示文本不足的问题,而且采用自适应方法,使整个过滤效果更加个性化。理论和实验均表明,该模型具有比较好的过滤效果,从速度和服务性能上得到了提高。

在模型的实现过程中,我们发现关键词的多义性影响了系统性能的提高,学习算法、相似度的衡量方法还存在一些不足之处。而且只依靠 WordNet 为知识来源也还是不足的,因此我们在下一步的工作中将增加常识库的建设,加强语义分析,提高文本表示的准确性。还有我们在今后的工作中,将学习一些新的机器学习算法,以使我们的自适应过程更加有效。

参考文献:

[1] D W Oard, et al. A Conceptual Framework for Text Filtering, University of Maryland[R]. Technical Report EE-TR-96-25CAR-TR-830CLIS-TR-96-02CS-TR-3643, 1996.

[2] 黄萱菁,等. 基于向量空间模型的文本过滤系统[C]. 中国中文信息学会二十周年学术会议, 2001. 12-13.

[3] 林鸿飞,姚天顺. 基于示例的中文文本过滤模型[J]. 大连理工大学学报, 2000, (5).

[4] 陈群秀. 一个在线语义词库:词网 WordNet[J]. 语言文字应用, 1998, (2).

[5] S Roberton, D Hull. The TREC-9 Filtering Track Final Report [C]. Proceeding of the Ninth Text Retrieval Conference (TREC-9), 2001.

[6] 战学钢,姚天顺. 基于汉语分析的中文分类方法[C]. 98 中文信息处理国际会议. 北京:清华大学出版社, 1998. 412-417.

[7] J Morris, G Hirst. Lexical Cohesion Computed by Thesaural relations as an Indicator of the Structure of Text[J]. Computational Linguistics, 1991, 17(1): 21-48.

[8] Stairmand M A. A Computational Analysis of Lexical Cohesion with Applications in Information Retrieval[D]. PhD Thesis, DeparUnent of Langu # Engneering. University of Manchester Institute of Science and Technology, 1996.

[9] Hirst G, D St-Onge. Lexical Chains Representations of Context for the Detection and Correction of Malapropisms[C]. Fellbaum, C. (ed.) WordNet. An Electronic Lexical Database, The MIT Press, 1998.

[10] Barzilay, Regina, et al. Using Lexical Chains for Text Summarization[C]. Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS '97), ACL Madrid, 1997.

作者简介:

尤文建(1977-),男,硕士研究生;李绍滋(1963-),男,副教授,"九三"社员,原机械部跨世纪学术骨干,研究方向为自然语言处理及智能信息检索、多媒体网络应用及计算机支持的协同工作(CSCW)等;李堂秋(1944-),男,教授,系主任,兼任福建省高校计算机等级考试委员会副主任,中国人工智能学会常务理事,自然语言理解与模式识别分会副会长,机器翻译分会副会长,厦门计算机学会理事长,中国中文信息处理学会计算语言学专委会委员等,主要研究方向为人工智能,特别是机器翻译等。