计算机科学2000Vol.27№.2

基于概念扩充的中文文本过滤模型*)

Concept Expansion and Chinese Text Filtering Model

林鸿飞　战学刚　姚天顺
(东北大学计算机科学研究所　沈阳110006)

Abstract The background and the future of text filtering are described in this paper, and a concept-based Chinese text filtering model is presented. The main idea of the model is shown as follows: Original keywords are given by users, then the concept expansion is automatically performed with the keywords to construct the user profiles. It is noted that user profiles consist of the subprofiles, and the ratio of sub-profiles matching and the similarity of sub-profiles are defined. As a result, they can weaken the Boolean constrains to ensure that the text can be matched while some sub-profiles don't match with it, and they can restrict the Vector Space Model to match more subprofiles. In addition, the mechanism of passage matching is applied to improve the efficiency of filtering model.

Keywords Text filtering, Boolean constrains, Vector space model, Concept expansion, Passage matching, User profiles, Fuzzy logic

1　前言

今天，以因特网为主体的信息高速公路仍在不断普及和发展，因特网上蕴涵的海量信息远远超过人们的想象，面对这样的信息汪洋大海，人们往往感到束手无策，无所适从，出现所谓的"信息过载"问题。如何帮助人们有效地选择和利用所感兴趣的信息，同时保证人们在信息选择方面的个人隐私权利?这已成为学术界和企业界所十分关注的焦点。因此，信息过滤技术应运而生。目前信息的表现形式大多数为文本，文本过滤技术仍得到迅速的发展，已成为文本处理的热点，也是计算语言学领域中新的增长点。在著名的 TREC(文本检索会议)中，过滤是路由任务(routing task)的重要子任务，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。

2　基于概念扩充的中文文本过滤模型

文本过滤的主要流程就是依据用户的信息需求(即用户模版)，在新文本流中查找与需求相匹配的文本。因此，文本过滤的关键任务，一是建立用户模版，二是文本与模版的匹配技术。

在建立用户信息需求模型(用户模版)时，一般由用户输入自己感兴趣的一些词，构成最初始的用户模版。然而，人们对同一事物其观察角度有所不同，所以对同一概念具有多种表达形式。另外，在文章撰写时因修辞的缘故，为了避免用词重复，常常出现同义替换现象。因此，须将用户的信息需求映射到概念一级以更加全面地反映用户的需求。但此时由于概念扩充会使过滤出的文本当中，真正相关的文本可能不多，而不相关的文本却过多，造成精度下降。为此要加以限制，尽量避免因扩充带来的噪音。

2.1　概念扩充

概念是人类对客观世界认识的结果，本质上是符号化的实体，它表示的是客观世界中的事物及其含义。每个概念可以由它的属性来描述，同时受到相关概念间的相互关系的制约，这些关系包括分类关系、构件关系、有序关系等。假定一个概念体系的所有概念都是有联系的，即存在一个共同的祖先，这样的概念间的关系是一个分类层次的偏序关系，可以形成一棵语义分类树。概念的层次不同，表明抽象程度不同。

在模型中，采用 CETRAN[1]的概念词典，其信息十分丰富，包括词条信息、词法分析和生成的信息、句法分析和生成的信息、语义分析的信息和概念处理信息。在词典中语义码相同的词为同义词，一组同义词为一个词群，整个词典约7万词，形成约4000个词群。

*)国家自然科学基金资助项目(编号:69675019)和国家教委博士点基金资助项目

假设用户提出的关键字集合为：$U=(t_1,t_2,\cdots,t_n)$，且关键字间的关系为"AND"。下面将其映射到概念空间的概念特征集 V，定义概念映射如下：

$$U=(t_1,t_2,\cdots,t_n)\xrightarrow{\Phi}(\Omega(t_1),\Omega(t_2),\cdots,\Omega(t_n))=V$$

其中 $\Omega(t_i)=((c_{i1},w_{i1}),(c_{i2},w_{i2}),\cdots,(c_{in_i},w_{in_i}))$，这里 $\{c_{i1},c_{i2},\cdots,c_{in_i}\}$ 是概念词典中与 t_i 具有相同语义码的概念集合，w_{ij} 是 c_{ij} 的权重，$w_{ij}=1$ 若 $t_i=c_{ij}$，此时称 w_{ij} 为 t_i 原始权重；$w_{ij}=\alpha(0<\alpha<1)$ 若 $t_i\neq c_{ij}$，此时称 w_{ij} 为 t_i 的扩充权重。如果在概念词典中不存在 t_i，则 $\Omega(t_i)=((t_i,1))$。每个 Ω 称为概念侧面，V 的所有侧面构成用户模版。

2.2 匹配算法

建立用户模版以后，如何依据模版过滤文本呢？这就是下面所要讨论的问题：计算文本与用户模版的相关程度的匹配算法。

若按照传统的矢量空间模型方法，计算文本特征向量与用户模版之间的内积，将其作为相似系数，公式如下：

$$Sim(U,V)=\frac{1}{(\|U\|\|V\|)^{1/2}}\sum_{i=1}^{n}u_iv_i \tag{1}$$

其中 U 为用户模版，V 为文本特征矢量。

依据选定的相关阈值，筛选符合条件的文本，结果往往产生大量匹配的文本，但真正相关的文本可能较少，或者排列的位置不当。这是因为在内积运算中，有些文本对于模版中的一个侧面或几个侧面具有较高的匹配程度，从而获得较大的相似系数，但对于其他侧面则可能完全不匹配，从而不能很好地全面满足用户的信息需求。

针对上述这样情况，在设计匹配算法时，不仅要考虑相似系数的大小，更重要的是要考虑文本对于用户模版的各个侧面的匹配程度，使之更加全面地满足用户的信息需求。鉴于这样的要求，本文选择布尔模型来加强模版侧面的匹配情况，同时用矢量空间模型计算文本与模版每个侧面的相似程度。

布尔模型：

$$(c_{11}\mathrm{OR}c_{12}\cdots\mathrm{OR}c_{1n_1})\mathrm{AND}(c_{21}\mathrm{OR}c_{22}\cdots\mathrm{OR}c_{2n_2})\cdots\mathrm{AND}(c_{m1}\mathrm{OR}c_{m2}\cdots\mathrm{OR}c_{mn_m}) \tag{2}$$

矢量模型：（为了叙述方便，将用户模版和文本特征矢量组织成相同结构）。用户模版为：

$$U=(0,0,\cdots,0,u_{11},u_{12},\cdots,u_{1n_1},u_{21},u_{22},\cdots,u_{2n_2},\cdots,u_{m1},u_{m2},\cdots,u_{mn_m})$$

文本的特征矢量：

$$V=(v_{01},v_{02},\cdots,v_{0n_0},v_{11},v_{12},\cdots,v_{1n_1},v_{21},v_{22},\cdots,v_{2n_2},\cdots,v_{m1},v_{m2},\cdots,v_{mn_m})$$

$$Sim(U,V)=\sum_{i=1}^{n}g(p_i,V)=\frac{1}{(\|U\|\|V\|)^{1/2}}\sum_{i=1}^{n}\sum_{j=1}^{s_i}u_{ij}v_{ij}$$

其中 $g(p_i,V)=\frac{1}{(\|U\|\|V\|)^{1/2}}\sum_{j=1}^{s_i}u_{ij}v_{ij}$，$i=1,2,\cdots,n$ 称做文本 T 的侧面 p_i 相似度，用来表示文本与用户模版侧面的相似程度。一个相关文本应与模版的各个侧面都具有很好的相似度，因此，对每个侧面相似度应有一定的要求。设布尔模型为：

$$p_1\mathrm{AND}p_2\cdots\mathrm{AND}p_n \tag{3}$$

其中 $p_i=(c_{i1}\mathrm{OR}c_{i2}\cdots\mathrm{OR}c_{in_i})$，$i=1,2,\cdots,n$，为用户模版的某一侧面。如果将 p_i 看作是一个模糊命题，则令 $\tilde{U}$ 为这样命题组成的集合，定义：

$$T:\tilde{U}\to[0,1],Q\mapsto T(Q),$$

此处 $T(Q)=g(Q,V)$ 为命题 Q 的真值，即选用侧面相似度作为命题 Q（侧面）的隶属度。按照模糊逻辑的运算法则，得到式(3)作为模糊命题的最终真值为 $T(P)=\min(T(p_1),T(p_2),\cdots,T(p_n))$。如果命题 $T(p_i)\geq\lambda(\lambda\in(0,1])$，$i=1,2,\cdots,n$，则称命题 $T(P)$ 为 λ-恒真。此时，定义用户模版的侧面匹配率如下：

$$\Psi(\lambda)=m/n$$

其中 n 是侧面数目，m 是作为模糊命题其真值不小于 λ 的侧面个数。它通常用百分数表示，如85%，即至少85%以上的侧面，它们与文本的相似度高于指定的阈值 λ。如此这样，可以根据需要，调整 λ 和匹配率，当少数侧面的相似度较小或者为零时，也有可能匹配成功，它兼顾了文本与模版的整体相似性和文本与各个侧面的匹配情况，对于传统的矢量空间模型，即公式(1)，没有考虑到与各个侧面的匹配问题，因此，可能会产生由于少数侧面的高频匹配，造成较高的相似度。而布尔模型，即公式(2)，当且仅当所有侧面的逻辑值都为真时，表达式才为真，即匹配成功。如果有一个侧面逻辑值为假，则匹配就失败。针对这两种问题，本文提出的过滤模型的匹配成功的条件是，不但文本与模版的相似度大于给定的阈值，而且侧面的匹配率也要达到给定的百分率，从而弥补了上述缺点。

2.3 段落匹配

近年的研究[6-7]显示，段落匹配机制明显改善了文本的检索效率。在文本过滤中，段落匹配也有改善精度的作用。但是，值得指出的是在文本中并不是所有的段落的重要性都相同，段落的位置以及对于文本主题的贡献不同，决定了段落的重要性。大量统计资料表明，首段和尾段都含有重要的主题信息，因此应给予较高的权重。尤其对于新闻语料，Searchable Lead[3]系统仅仅从文章开头部分抽取给定长度的一部分形成摘要，就达到87%—96%的可接受率。所以应将首段、尾段和其它段在权重方面有所区分。

本文采用段落匹配的主要目的在于提高查准率，

降低存在的"噪音"，从而使在某个或某些段落上具有较好相似性的文本应优于那些匹配特征比较分散的文本。

将上述的匹配算法应用到文本的各个段落，相当于将每个段落看成是独立的文本，按照相似系数从大到小排列输出指定数目的段落，然后计算对应文本的最终权重如下：

$$w(T)=\sum_{i=1}^{m}h(p_i)Sim(p_i)$$

其中 m 表示参与权重运算的段落个数，一般 $m=3$，即在输出的段落中，选取文本 T 权重最大的前三个段落参加计算，可能有些文本被输出的段落数小于3个。$h(p_i)$表示段落的权重，若 p_i 为首段，则 $h(p_i)=1.0$；若 p_i 为尾段，则 $h(p_i)=0.95$；若 p_i 为其它段，则 $h(p_i)=0.80$，$Sim(p_i)$是段落与用户模版的相似度。过滤模型的输出结果依照文本的最终权重进行。

值得指出的是段落匹配的另一个功能，在对冗长文本的过滤过程中，直接给出与模版想匹配的段落，略去无关的段落。尤其对于类似于百科全书的文本，如各类百科全书、条文、规定和政策性文献等，人们感兴趣的是非常具体的条款，而不是全文。

2.4 用户模版与过滤流程

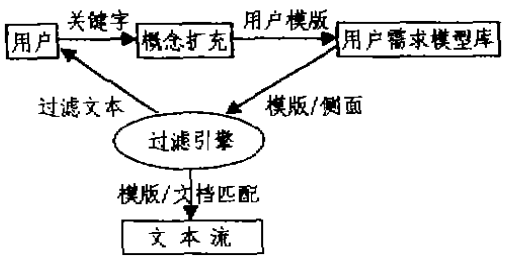

图1 基于概念的中文文本过滤模型

基于概念的中文文本过滤模型采用倒排索引方式表示用户模版，每个模版包括两部分信息，一是组成模版的各个侧面，而侧面是概念扩充后的相应概念集合；二是模版的各类控制参数，包括总体相似度阈值、侧面相似度阈值和侧面匹配率。

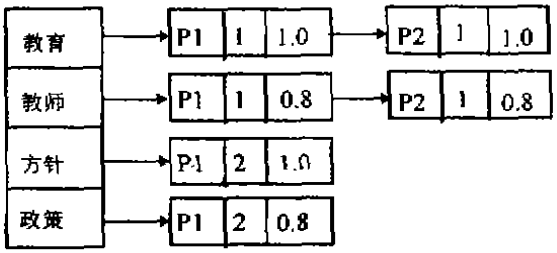

图2 概念-模版-侧面索引表

	总体阈值	侧面阈值	侧面匹配率
P1	0.30	0.10	0.50
P2	0.40	0.15	0.75

图3 模版参数表

概念-模版-侧面索引表(图2)是模版的物理表示，它是以概念为索引的链表，第一个域存储模版的编号，如 $P_1,P_2,\cdots,P_n$；第二域存储模版的侧面号，如1，2，…，m；第三个域存储概念在侧面中的权重。模版参数表(图3)中总体阈值表示文本与模版匹配时的相似系数应大于或等于该值；侧面阈值表示文本与模版侧面的相似度应大于或等于该值；侧面匹配率表示侧面相似度大于或等于该值的侧面个数应大于或等于该值，此时要求匹配的侧面个数至少为1。

例如：用户给出的关键字为"教育、方针"，则经过概念扩充，形成的用户模版为 P1＝(教育 OR 教师) AND(方针 OR 政策)，其中侧面1为(〈教育，1.0〉，〈教师，0.8〉)，侧面2(〈方针，1.0〉，〈政策，0.8〉)。计算文本与模版的侧面相似度，要求50%以上的侧面相似度大于等于侧面阈值，这里共有两个侧面则至少有一个侧面符合要求。假设文本为"教育改革的进展"，模版 P1 为(〈教育，1.0〉，〈教师，0.8〉，〈方针，1.0〉，〈政策，0.8〉)，侧面1的相似度为0.32，侧面2的相似度为0，总体相似度也为0.32，因此可以认为它与模版匹配。(以上概念扩充是部分扩充，仅供举例)。

3 实验结果

我们测试的语料为来自《人民日报》1994年共8000余篇文章，采用 CETRAN 的概念词典，它参考梅家驹先生《同义词林》，结合其它语义分类体系，构造而成。共构造10个用户信息需求模型，即用户模版，侧面匹配率为75%，侧面相似度阈值为0.05，总体相似度阈值为0.25，将段落看成独立的文本，抽取前800段落，计算文本的最终权重，对于用户模版中的各个概念其原始权重为1，扩充权重为0.80。

实验结果表明本文提出的过滤方法，与单纯矢量

(下转第82页)

```
member_of_project(A,B):-research_project(A),
    person(B),link_to(C,A,D),link_to(E,D,B),
    neighborhood_word_people(C)
department_of_person(A,B):-person(A),depart-
    ment(B),link_to(C,D,A),link_to(E,F,D),
    link_to(G,B,F),neighborhood_word_graduate
    (E)
```

3.3 文本段的抽取

有时候，要抽取的信息不能用 Web 页面或页面之间的关系来表示，而仅仅是嵌在页面中的文本段。同样是基于 FOIL，WebKB 开发了一种信息抽取的学习算法：SRV[4]，它是采用爬山法的一阶学习器，算法的输入是一些页面，页面中所要抽取信息的实例已经被标记出来。算法的输出是一些用于信息抽取的规则，而信息抽取的过程就是检查任何长度的文本，试图找到能够和规则匹配的文本片段。抽取出来的规则如下：

```
ownername(Fragment):-
    some(Fragment,B,[],in_title,true),
    length(Fragment,<,3),
    some(Fragemnt,B[pre_token],word,"GMT"),
    some(Fragment,A,[],longp,true),
    some(Fragment,B,[],word,unknown),
    some(Fragment,B,[],quadrupletonp,false).
```

其中，Fragment 是要匹配的文本段，length(Fragment，Relop，N)用于限制抽取信息中词的个数，其中 Relop 是比较运算符，N 是数字。

some(Fragment，Var，Path，Attr，Value)用于检验 Fragment 中词的特性，其中，Var 是当前词单元，Path 是被检验单元的路径(和本单元的关系)，Attr 是要检验的属性，Value 是属性的值。

下面是和以上规则匹配的万维网页面，其中 Bruce Randall Donald被抽取出来，被认为是本页面的属主。

```
LAST-MODIFIED: WEDNSDAY, 26-JUN-96 01:37:46
    GMT
<TITLE>BRUCE RANDALL DONALD</TITLE>
<H1>
<IMG
    SRC = "FTP: // FTP. CS. CORNELL. EDU/PUB/
        BRD/IMAGES/BRD. GIF">
<P>
BRUCE RANDALL DONALD<BR>
```

4. 各种挖掘方法的比较

ParaSite 的方法比较简单，精度不高，而且发现的知识十分有限；DRS 方法在专业范围内可用性比较好，也有一定的学习功能，但应用范围比较有限，一般只用于某个特定类别的信息搜索。WebKB 中提供了多种信息挖掘的方法，功能很强，既可以适应多个特定的类别，理论上也可以用于挖掘通用的信息，但它需要有一定规模的训练文本，而且，随着信息类型的通用化，训练文本的规模和质量也要随着提高，因此实现的难度也随着变大。这些方法，各有其特点，在构造万维网挖掘系统时，要综合使用这些方法。一般地说，ParaSite 中的启发性规则对大部分的系统都有所帮助，也容易实现，DRS 则针对比较专门的信息类型，而 WebKB 可以适应规模较大的系统。不同的系统，可以针对各自的特点，对这些方法进行剪裁。例如，在我们开发的 PersonIndex[6,7]中，使用了 ParaSite 中的一些启发性规则，以及 DRS 中的 Email 过滤和 URL 模式。由于 WebKB 中通用的训练文本很难得到，我们就使用专业的语料库来代替，如人名语料库，机构名语料库等来做局部的训练，也取得了很好的效果。

主要参考文献

1 Spertus E. ParaSite: Mining Structural Information on the Web. In: Proc. of the 6th World Wide Web Conf. 1997

2 Shakes J, Langheinrich M, Etzioni O. Dynamic Reference Sifting: A Case Study in The Homepage Domain, Proc. of WWW6

3 Craven M. Learning to extract symbolic knowledge from World Wide Web: [Technical report]. CMU CS Dept, 1998

4 Lewis D. Training algorithms for linear text classifiers. In: Proc. of the 19th Annual Int. ACM SIGIR Conf. 1996

5 Quinlan J R. FOIL: A midterm report. In: Proc. of the 12'h European Conf. On Machine Learning, 1993

6 Shen Dayang, YuBin, Lin zuoquan. WebIndex, the Intranet Information Collecting Agent. In: Proc. of APWeb98, Beijing, 1998

7 沈达阳，孙茂松. Internet 中文个人信息搜索，中文信息学报(已录用)，1999

(上接第90页)

空间模型相比平均精度有明显改善。前者为37%，后者为51%，增加14%。

参 考 文 献

1 姚天顺，等. 自然语言理解. 清华大学出版社，1995

2 吴立德，等. 大规模中文文本处理. 复旦大学出版社，1997

3 刘开瑛，等. 中文文本中抽取特征信息的区域与技术. 中文信息学报，1998，12(2)

4 Yan T W, Garcia-Molina H. SIFT-A Tool for Wide-Area Information Dissemination. In: Proc. of the 1995 USENIX Technical Conf.

5 Yan T W, Garcia-Molina H. Distributed selective dissemination of information. In: Proc. of the Third Intl. Conf. on Parallel and Distributed Information System. 1994. 89～98

6 Callan J P. Passage-Level Evidence in Document Retrieval. In: Proc. SIGIR' 94. 1994

7 Salton G, et al. Automatic Structuring and Retrieval of Large Text Files. Comm. of the ACM, 1994, 37(2): 97～108

8 Buckley C, et al. Automatic Query Expansion using SMART: TREC-3. In: Proc. of the Third Text Retrieval Conference. 1995