可分义原向量空间中的跨语种文本过滤模型

苏伟峰　李绍滋　李堂秋　尤文建
厦门大学计算机系
E-mail: szlig@xmu.edu.cn

摘要： 本文介绍一个可以从中文或英文大量的信息中过滤出用户的兴趣所在的文档的模型，该模型采用向量空间的方式，从用户提供的文档中学习，用可分义原向量空间的一个向量来表示用户所感兴趣的文本，然后把需要处理的文本也表示成一个可分义原空间中的一个向量，利用两个向量之间的夹角来判断两个向量的相似度从而决定是否将该文本呈现给用户。实验证明，这是一个比较好的过滤方法。

关键词： 可分义原 ，向量空间，用户模板，文本表示

A Cross-lingual Text Filtering Module In Classifiable Sememes Vector Space

Su Weifeng　Li Shaozi　Li Tangqiu　You Wenjian
Department of Computer Science, Xiamen University , Xiamen，361005
E-mail: szlig@xmu.edu.cn

Abstract: This paper decribes a module that sift through large number of source text in Chinese or in English which may be of the user's interest. It is an approach that learns the model of the user's preference, filters the information, and notifies the user when relevant information available. The user's model is represented as a vector in the vector space of classfiable sememes. The document to be filtered is also represented as a vector. The relevance of the text to the user can be measured by using the cosine angle between the user vector and the text vector. Experiments show that it is a good idea.

Keyword: Classfiable Sememe, Vector Space, User Model, Text Representation

1. 引　言

随着因特网和其它在线信息资源的迅猛发展，大量的信息朝人们涌来，据统计现在全球大约有 80 亿页网页，而且这个数字还以每 3-5 个月翻一翻的速度上涨，显然在这信息海洋中仅靠原来的那种手工作坊式的方式来搜查所要的信息效率就太低了，人们希望能够自动分挑出有用的文本信息的机器的出现。

文本过滤是自动分挑出有用的文本的一种很重要的方法。文本过滤是指从大量的源

信息中过滤出那些最符合用户需求的信息传送给用户，而跨语种文本过滤是指源信息中包含多种语言（比如英语、汉语等），或者某个文本中就含有多种语言，从中过滤出用户所需要的文本，过滤出的文本可能也是多种语言的。在没有国界的因特网上，跨语种过滤出所要的信息就显得更为重要。在把大量的信息送给用户之前过滤掉那些用户不感兴趣的东西，这比在有条件后，翻译成某种语言过后再进行过滤更能省掉用户大量的精力和时间，跨语种过滤系统对于那些对需要这一语种的信息的而又对该语言掌握的不好用户特别重要。

在跨语种文本过滤方面，人们已经摸索出了许多方法来实现不同语种之间的相互转换形式。最初人们是提出一种基于控制词汇的方法[1]，即把文本表示成一些固定的词，用户的需求也表示成这些固定词汇，然后进行匹配。这个方法最大的缺陷是词汇必须在可管理的范围之内，而一旦词汇超出可管理的范围，则其召回率和精确率则迅速下降，而且如何把文本表示为词汇目前也没有一个很好的方法。

接着又有人提出基于字典的方法[2]，就是编辑一本多语字典把某种语言的文本表现形式通过翻译表示成另一种语言的表现形式，从而使那些单种语言上的文本过滤技术可以应用于多语言的文本过滤，这个方法理论上是有可能的，但是有两个方面的原因却限制了它的应用。首先是一词多义的现象，在翻译中一个词可能翻译成几个意思，若几个意思全都采用则大大降低了精确率，若采用某一个意思，则有可能降低召回率，或者根本就选择错误而导致召回率极低。第二是一义多词的现象，由于不同的作者可能用不同的词来表达同一个意思而导致召回率下降。

本文提出一种新的思路，它不从词这一级来分析概念，而把词所包含的概念进行分解，再进行分析，类似的思想在其他学科用过，比如我们分析某种物质时，常将其分解至其原子或分子水平，再根据原子或分子的特性从而得到该物质的特性。

2. 过滤模型的系统结构

我们采用的技术主要是向量空间模型，即把用户模板和文本均表示成向量空间中的向量，向量空间的优点是将文本内容转换成易为数学处理的向量方式，使得各种相似运算和排序成为可能。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，取得了良好效果。本文所提出的基于向量空间的文本过滤模型可以用于对中文和英文的文本进行过滤。其基本思想是首先利用用户所提供的材料来获取用户的模板，然后利用用户模板来判断某一文件是否与用户模板相近。

该模型的体系结构如图 1 所示。

该模型采用董振东先生所研制的《知网》[3]，该系统带有 53000 个中文词组和 57000 英语单词。《知网》是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库，《知网》采用义原来表示概念，义原是最基本的、不易于再分割的意义的最小单位，例如："人"虽然是一个非常复杂的概念，它可以是多种属性的集合体，但我们也可以把它看作为一个义原。我们设想所有的概念都可以分解成各种各样的义原。董振东先生提取出了 800 多个义原，并用

它们的组合来表示世上所有的概念，比如它是这样注释：“扭亏为盈”

DEF=alter|改变,StateIni=InDebt|亏损,StateFin=earn|赚。

即是指“扭亏为盈”是一种“改变”，其起始状态是“亏损”，最终状态是“赚”。

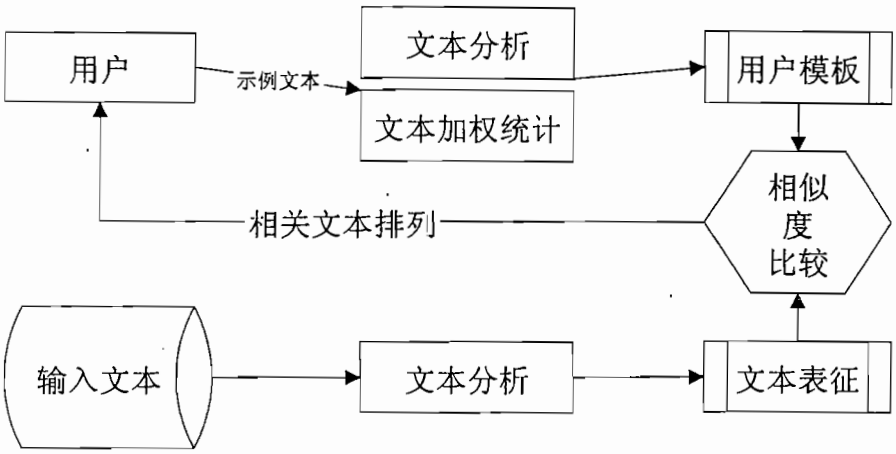

图 1 ：可分义原向量空间中的跨语种文本过滤模型的体系结构

把概念分解成为义原可以极大限度地解决一义多词的问题，比如“电脑”、“计算机”、“computer”这三个词，在《知网》里均定义为“computer|电脑”，这样我们就可以把它们视为概念等同的三个词语。其实从这个层面上来理解，我们可以把某个词的中英文意思同样看作是一个一义多词的一种形式，这样只要解决好了排岐的问题，我们并不需要特殊处理就可以解决跨语种的问题，因而我们的方法可以说是一种与语言无关的方法。

《知网》中的许多资源对于解决一词多义的问题也是颇为有效的，这将在下文讲到。我们把义原继续分为两类：可分义原和不可分义原。不可分义原是指那些表示的意思比较过于常见，没法用来指出该概念一些特有的性质的义原，而可分义原则是指那些能表示该概念的重要的可与别的概念区分的开的义原。若我们不排除掉不可分义原，则由于不可分义原的较高的出现频率，就有可能误导我们。可分义原在本系统中占着重要的地位。

本文下面所用到的技术对于中文文档和英文文档同时适用，若是有不同的地方则会分别指出。

3. 过滤模型的设计和实现

3.1 文本表示方法

我们采用的技术是向量空间模型，文本表示为向量空间中的一个向量。向量空间表示为 $\vec{D}$ ，而每一个分量 d_i 是知网中的一个可分义原，那文本就表示成向量 $\vec{V}$ ，其分量 v_i 为对应于 d_i 的值，若文本中没有包含 d_i ，则 v_i =0。

然而并非文件当中所有的词都用于构造文本向量，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向量。我们可以采用统计的方法来决定哪些词汇是关键词汇，还有，由于词汇的歧义，我们也要作一定程度上的排歧。文本表示方法可归纳如下：

1） **文本预处理**。包括判断文本是中文还是英文的，或是中英文混杂。如果是中文文本，我们还需要将其进行单词切分。当然，对于中英文混杂的文本，我们需将其中文部份进行单词切分。

2） **词性标注**。对文本中的单词进行词性标注。

3） **关键词抽取**。在英语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，在中文文本中去除所有的虚词，这样在文本就剩下主要的词像名词、动词、形容词和副词，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给赋予不同的权值来表示它们不同的重要性，一般而言，名词要赋以最大的权值。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。我们也可以设一个阀值，把那些出现频率低于该频率的词去除。

4） **关键词概念排歧**。过多的歧义会妨害我们向量表示该文本的效果，尤其当某个词在该文本当中占有比较重要的地位时。排歧的基本思想是根据上下文词的义原对该词为某一意思进行概率统计。其主要思想是：在一篇文章当中，某个词会对上下文的用词产行影响，通过上下文可以判定某个词的意思从而进行排歧，在本模型下，考虑着重考虑其上下文当中其他的关键词的义原与该词的义原有无以下情况：

a.有相同可分义原，

b.材料-成品关系，

c.施事/经验者/关系主体-事件关系

d.受事/内容/领属物等-事件关系

e.工具-事件关系

f.场所-事件关系

g.时间-事件关系

h.事件-角色关系

i.相关关系

如果其上下文当中某个关键词当中有个可分义原与该词的某个意思某一可分有上述关系，则增加该意思的权重。

在 **W** 中，对某个词 *w*，在以其为中心的窗口宽度为 n 的字符串表示为：

$$w_1 w_2 \ldots w_{n/2} w w_{n/2+1} \ldots w_{n-1}$$

对于 *w* 在知网中的每一个意思，赋予一个初权 k，调节词 *w* 每一个意思的权值的方法的伪代码算法 1 所示：

算法 1 ：词的义原的权值的调节

W_I—窗口中除去 *w* 的第 I 个词

S_{IJ}—窗口中除去 *w* 的第 I 词的第 J 个意思

CS_{IJK}--窗口中除去 *w* 的第 I 词的第 J 个意思的第 K 个可分义原

WS_J—词 *w* 的第 J 个意思

WCS_{JK}--词 *w* 的第 J 个意思的第 K 个可分义原

Weight(WCS_J)—词 *w* 的第 J 个意思的权值

```
FOR I=1 TO n-1    //对于窗口中除了 w 外的每一个词
  FOR J=1 TO  (W_I 的意思数目)
    FOR K=1 TO  ( S_IJ 的可分义原数目)
      FOR M=1 TO  (词 w 的意思数目)
        FOR O=1 TO  (WS_J 的可分义原数目)
          IF CS_IJK 与 WCS_JK 有上述关系 THEN Weight(WS_J)= Weight(WS_J)
                                              +1
          ENDIF
        ENDFOR
      ENDFOR
    ENDFOR
  ENDFOR
ENDFOR
```

由此，词语的那些与上下文相关的意思都通过增加权值而得到加强，当然我们还要对此进行归一化处理，基归一化的公式如下所示：

$$wt(WS_i) = \frac{Weight(WS_i)}{\sum_i Weight(WS_i)}$$

其中 *i* 是该词的意思的序号。

5） 文本表示成一向量。在经过了关键词提取和排岐之后，我们把这些关键词根据其义原权值按照知网里的单词定义分解成为义原，并在去除了不可分义原之后，我们采用算法 2 中的方法计算各可分义原，文件就表示成了可分义原空间中的一个向量。

算法 2 把一个文件表示成可分义原空间的一个向量算法

V_K— 向量中的分量的值

SM_{IJK}—第 I 个关键词第 J 个意思的第 K 个可分义原

Weightof(SM)—某个可分义原的标量值

$wt(S_{IJ})$－第 I 个关键词第 J 个意思的权值

```
给向量的每个分量值赋初值 0
FOR I:=1 TO  (关键词的数目)
  FOR J=1 TO  (第 I 个关键词的意思总数)
    FOR K=1 TO (第 I 个关键词第 J 个意思)
      Weightof(SM_K)= Weightof(SM_K)+wt(S_IJ)
    ENDFOR
  ENDFOR
ENDFOR
```

3.2 用户模板表示

和把文本表示成为一个向量一样，我们把用户模板也表示成为可分义原向量空间中的一个向量。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，可以是英文材料，或可以是中文材料，但是这些材料要属于同一个类别之内的材料，如果用户提供的是摘要或文章，则需要象在第 2 部份中的那样把他们表示为向量空间中的向量。如果用户提供的是关键词，则把根据这些词如第 2 部份那样进行排岐，再把它表示成可分向量空间的一个向量，所有这些向量的均值形成了用户模板的向量。当然我们可以给用户给其所给的文本赋以不同的权值以表示其不同的重要性。在本模型中，在求用户模板之前，我们对向量进行归一化处理，同时文章的长度也作为考虑因素之一。我们按如下公式给每一文本赋予不同的权值：

$$weight(\text{text})=log(N)$$

其中 N 表示文本关键词的数目，我们还设置了一个阀值，所有小于该阀值的分量均赋值为 0，以此实现降噪处理。计算用户向量具体如算法 3 所示。

算法 3：用户向量计算。

Wight($TEXT_I$)—用于构造用户向量的第 I 个文本的权值

S_{IJ}—用于构造用户向量的第 I 个文件的向量的第 J 个分量的值

S_J—用户的第 J 个分量

```
对用于构造用户向量的每个文件向量进行归一化处理。
给用户向量的每个分量值赋初值 0
FOR I=1 TO  (用于构造用户向量的文本数)
  FOR J=1 TO  (可分义原的数目)
    S_J= S_J + Wight(TEXT_I)* S_IJ
  ENDFOR
ENDFOR
FOR I=1 TO  (可分义原的数目)
  IF S_J<VALVE THEN S_J =0
  ENDIF
ENDFOR
```

3.3 文本过滤

至此为此，文本和用户的需求者已表示成一个向量，文本与用户需求的相关度可以通过根据这两个向量之间的夹角的余弦值而得：

$$\cos(a) = \frac{(V_{user}, V_{text})}{|V_{user}||V_{text}|}$$

其中(V_{user}, V_{text})是指用户向量和文本向量的内积，而$|V_{user}|$和$|V_{text}|$分别是指用户向量和文本向量的标量。

在所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，也可以设一阀值 t，当某文本与用户需求的相关度大于 t 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。我们可以把用户的回馈考虑进去，若用户认为几乎所有我们所过滤出的文件都是他所感兴趣的，则我们可调低 t 值，反过来，若有很多文本不符合用户的兴趣，则我们调高 t 值。

4. 过滤模型的实验结果及实验分析

我们获得了八个用户的实验数据，这八个用户都提供了他所感兴趣的内容相近的中英文文本各 40 篇作为相关文本，另外提供 200 篇其它内容的文本作为干扰文本，其中中英文各 100 篇，对于每个用户，我们使用从其所提供的相关文本随机抽取中英文文本各 20 篇构造其用户模板，其余的相关文本与干扰文本混杂一起构成了测试集，我们就想从其中过滤出那些相关文本。

我们使用了两个参数来评价我们的模型：召回率和精确率。召回率是指我们过滤出的相关文本占所有相关文本的比率，精确率是指在我们所有过滤出的文本当中，相关文本所占的比率，一般而言，召回率上升，则精确率会下降，而精确率上升，则召回率会下降。

表 5 就是我们实验的结果，结果表明用该方法进行过滤的方法效果非常好，精确率很高，在实际应用当中，我们还可以把用户反馈的情况考虑进去，形成可根据用户的兴趣改变而把改变用户模板向量从而改变选择的文本的自适应系统。

		User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	Average
召回率(%)	English	87.5	90	90	90	85	85	90	87.5	87.5
	Chinese	85	85	87.5	87.5	85	85	90	87.5	86.56
精确率(%)	English	85	90	83	81	80.5	80.5	81	83	83
	Chinese	83	83	81	83	78.4	80.5	83	81	81.6

表 5：使用该方法的八个用户的召回率和精确率

我们可以从以下几方面来分析这个过滤模型产生较好结果的原因：

1. **低维分析空间**：所有的概念都被分解成义原，只须在可分义原空间中计算相似程度，这样我们就只要计算 400 个左右的可分义原而不是 100000 个左右的中英文单词，如此降低维数可极大地提高召回率，还有，可以降低计算复杂度。
2. **相关分量值较大**：比如在一篇病人上医院去看病的文本里，可能会会出现许多类似“病人”、“医生”、“医院”、“治疗”等词组，这些词都包含有“医治”等义原，从而使“医治”这个义原分量的值比较大，这样就能突出本文的所要讲述的内容主要是关于医疗这一方面的，这有助于提高精确率与召有率。
3. **干扰项较少**：经过了关键词提取、词语排岐和不可分义原的去除后，所剩下的义原大多与文本有重要的联系，而与文本相关度较少的其他分量的值相比之下明显较小。

5. 结束语

从网络信息服务需求出发，我们认为有必要对信息源的信息进行过滤。本文提出了一个在可分义原空间中采用向量空间模型的方法进行文本过滤的模型，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达到了较好程度。

在模型的实现过程中，我们发现关键词抽取精度的提高，在相当程度上可提高过滤模型性能。所以，提高关键词抽取的正确率，并将抽取领域扩展为名词、动词、名词短语及动词短语的综合，成为我们将来的研究。

参考文献

[1]TRANSLIB. Advanced Tools for Accessing Multilingual Library Catalogues. Technical Report,Deleveralbe D.1.4:Evaluation of Tools.Knowledge S.A.,June 1995

[2]L.Ballesteros,W.B. Croft. Dictionary-based methods for cross-lingual information retieval, Proc. Of the 7 th Int. DEXA Conference on Database and Expert Systems Applications ,1996.

[3]董振东，董强.， 知网, http://www.keenage.com/html/index.html.

[4]Douglas W.Oard,Gary Marchionini,A Conceptual Framework for Text Filtering, http://citeseer.nj.nec.com

[5]张月杰，姚天顺. 基于特征相关性的汉语文本自动分类模型的研究. 小型微型计算机系统，1998 年第 8 期。

[6] 苏伟峰、李绍滋、李堂秋. 一个基于概念的中文文本分类模型. 计算机工程与应用. 待发表