文本褒贬倾向判定系统的研究

孟凡博，蔡莲红，陈 斌，吴 鹏

（清华大学计算机系北京 100084）

E-mail ：mfb03@mails.tsinghua.edu.cn

摘 要：为了满足当今对评论性信息进行分析的需要，本文设计并实现了一个基于关键词模板的文本褒贬倾向判定系统。本系统定义了关键词类别、建立了关键词库、关键词模板库，并设计了模板匹配算法和文本褒贬倾向值算法，对测试文本进行关键词及模板匹配进而判断测试文本的褒贬倾向。本文还对文本褒贬倾向判定系统进行了测试、分析，测试及分析结果证明在语料充足的条件下，本系统可以有效的判断文本的褒贬倾向。

关键词：关键词分类，关键词匹配，模板匹配

中图分类号：TP301 **文献标识码：**A **文章编号：**0800109

Research on the recognition of text valence

Meng Fan-bo，Cai Lian-hong，Chen Bin，Wu Peng

（Department of Computer Science and Technology, Beijing 100084, China）

Abstract: Today more and more information is available, and more and more people are focus on how to make use of the information. The analysis of commentary information has attracted close attention of the scientists. Thus this paper proposed a recognition system of text valence based on key word template. This system defined key word classification and key word templates, built key word libraries and key word template libraries. This paper proposed template matching arithmetic, text valence value arithmetic. This system did key words matching and templates matching. Then this system calculated the valence of the testing-text. Besides this paper did a test on the recognition system. The analysis result proved this method was effective, given enough training corpus.

Key words: Key word classification, Key word matching, Template matching

1 引言

随着计算机软硬件技术和网络技术的飞速发展，人类文明也逐渐从工业化社会迈入了信息化社会。在信息资源日益丰富的今天，一方面丰富的信息为人们生活带来便利，另一方面庞杂的信息也常常使人无从下手。于是，如何快速有效地利用信息成为人们关注的焦点。在各种各样的信息中，评论性信息又是十分经典的一类。一方面这类信息具体内容涵盖了新闻、舆论、产品、技术、生活等许多方面。另一方面这类信息反映了人们对事物的主观看法。因此，如何更有效的对评论性信息进行处理、分类、提取有用信息得到了持续而广泛的关注。

目前，文本信息挖掘的研究工作，有些已取得很多成果，并付诸应用，比如主题分类、摘要提取等。但有些课题的研究却刚刚起步，如文本褒贬、情感倾向的分析，主要还是停留在词汇情感倾向的层次上[1][2]。要判定文本的褒贬倾向，要对文本进行句法分析、语义分析，以及褒贬倾向判决。

当前，许多学者对汉语自动分词方法进行了研究并取得了很好的成果[3][4]。本系统的前端句法分析部分使用了中国科学院计算中心研究所的汉语词法分析系统ICTCLAS[5]，其分词速度及准确率都有令人满意的表现。

在汉语词汇语义倾向判定方面，Turney 的研究，即使用 SO-PMI 的语义分析的方法需要计算并比较目标词汇与两个语义倾向相反的基准词汇的词语间的语义倾向[2]，该方法需要有极为丰富的语料库为基础，因此很难应用于实际的

收稿日期：2008-- 基金项目：本工作由国家自然科学基金China National Natural Science Foundation(60433030,60418012)资助 作者简介：**孟凡博**，博士研究生，主要研究方向为情感语音；**蔡莲红**，博士生导师，研究方向为主要研究方向为语音合成，基于内容语音检索，语音可视化；陈斌，硕士研究生；吴鹏，硕士研究生。

系统中。另一种计算词汇语义倾向的方法是利用基于《知网》的词汇语义相似度计算的方法[6][7],。但是只有语义的相似度还不够，与之相关的副词、否定词还会影响到语义的倾向。

本文设计了基于关键词模板匹配的文本褒贬倾向的判定系统，系统结构如图 1。本系统在得到原始语料文本并对其进行基本的分词之后，进行关键词匹配形成待匹配的关键词序列，之后通过模板库进行模板匹配，若模板匹配失败则对待匹配的关键词序列进行模板模糊匹配的特殊处理，最终计算文本的褒贬倾向。其中分词采用中国科学院计算中心研究所的汉语词法分析系统 ICTCLAS[5]。词汇语义相似度参考《知网》。

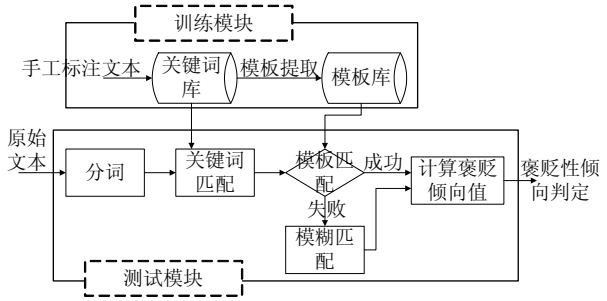

图 1 系统结构示意图

Fig. 1 System structure schema

本文将在第二节介绍系统中关键词词库的构建以及关键词的匹配，第三节介绍模板的提取与匹配。第四节说明褒贬倾向的计算。最后，本文对系统进行了集内、集外两项测试，并对测试结果进行了分析。

2 关键词的分类、标注及匹配

所谓关键词，是指对文本语句的褒贬倾向有决定作用的词汇。关键词的提取与匹配是系统对原始语料文本的前端处理模块，是对原始语料文本去芜存菁的过程。

2.1 关键词的分类及标注

ICTCLAS 是一个高准确度、快速的分词系统，它得到的分词结果词性分类超过 20 种。考虑到对于文本的褒贬倾向分析，减少匹配模板的数量以及缩小模板库，本文重新标注了词性，主要为名词、动词、形容词、连词、介词、副词和其它七种。每类词在汉语语句中的位置以及所起的作用一般比较固定，利于模板的提取和匹配。

在词性标注基础上，本文又设计了三种标注，分别为类别、褒贬性和程度。类别根据每个词语对文本褒贬性影响的不同而将词语分类。目前本系统将词语分为五类：

1、直接能表达出褒贬倾向的词汇（包括一些名词、形容词、副词和动词，如精彩，荒诞）

2、表示程度的副词（很，非常）

3、否定词（不，没有）

4、表示转折的连词（但是，却）

5、某些合成词，即按分词的结果拆开单独看不带情感，但是整体带有情感倾向的词组。例如：创世纪，分词系统将它分成两个词，这两个词分别出现并不带有褒贬倾向，而当同时出现时，则带有一定的褒义倾向。(创世纪，载入史册)

本系统仅标注这五种类别的词语，其它类别的词语系统认为对文本褒贬性没有影响，所以不进行标注。类别随着系统的改进，可以继续添加。

褒贬性指的是当前词语所具有的褒贬性，这项标注仅对类别为 1 和 5 的词语有意义。程度指的是当前词语所具有的褒贬性的程度或对于褒贬程度的影响，这项标注对于类别为 1、2、5 的词语有意义。

另外，标注的格式设计也是一个很重要的方面，关系到系统词库建立模块的效率。首先标注应该拥有统一的长度，这样方便系统的处理，其次标注要能够体现出各种不同词类之间的区别，最后标注还应该提供一个有效的方法来表达第 5 类的合成词。本文设计的标注格式为：/[a,c,d,n,v,p,i]（词性）[1,2,3,4,5]（类别）[+,-,#]（褒贬性）[数字]（程度）。表 1 为一句话分词后并添加标注的例子，其中，很/d2#2 表示程度副词，本身不具有褒贬性，对于褒贬性的影响因子为 2。而精彩/a1+2 则表示形容词，具有褒义情感，情感程度为 2。

表 1 关键词标注

Table 1 Key word annotation

标注步骤	标注结果
原始文本	这部电影很精彩。
分词结果	这/r 部/q 电影/n 很/d2 精彩/a1 。/w
标注结果	这/r 部/q 电影/n 很/d2#2 精彩/a1+2 。/w
关键词序列	很/d2#2 精彩/a1+2

2.2 关键词的匹配

首先，在关键词的训练阶段，本文从网上选取了 33 篇电影《变形金刚》的评论，从中人工筛选带有褒贬倾向的语料，并人工对它们进行标注，从而得到用于训练的语料。系统针对五种类别的词语分别建立关键词库，词库中完整保存 33 篇训练语料中的词语以及它的标注。目前关键词库的规模及示例如表 2。

表 2 关键词库规模

Table 1 Size of key word dictionary

关键词类	规模（个）	示例
1	235	美好/a1+2
2	88	都/d2#3
3	41	毫无/v3#2
4	15	尽管/c4#2
5	18	创世纪/n5+3

其次，在关键词的匹配阶段将原始文本分词之后，根据词性的不同到不同的关键词库中匹配，如果匹配上，则补上训练时保存的标注。这里需要对类别为 5 的词语特殊处理。

对于合成词的匹配，本文设计了下面的的方法来解决：

对于一个合成词，设它被分词程序分为 k 个词：C1，C2，……，Ck，把针对第 5 类词的词库分为两个部分，一个部分记录这个合成词的前缀，即 C1；另一个部分记录则整个完整的合成词。匹配的过程如下：

1、当匹配某一个词 C 的时候，我们首先在在前缀库中查找，看它是否存在与前缀库之中。如果能在前缀库之中找

到，那么转 2)，否则转 4)。

2、将 C 后面的词拼接到 C 的后面，形成一个新的词 C’，判断 C’是否存在词库中，如果找到，转 3)，否则令 C = C’，转 2)。直到 C’的长度超过词库中以 C 为前缀的最长词的长度，转 4)。

3、匹配合成词成功，直接跳过整个 C’，匹配 C’后面的词。

4、合成词匹配失败，到其余 4 类词库之中匹配 C。

利用这样的算法，可以成功地匹配存在的合成词，又能在匹配不成功的情况下正常匹配词库之中其他 4 类词。

3　模板的提取与匹配

模板的提取与匹配是从关键词序列中抽取共性的过程。在处理模板的过程中，一方面模板不能过于抽象，使得模板匹配时准确率下降；另一方面模板也不能过于具体，从而不但增多了模板的数量，模板也失去了其本来的意义。

3.1 模板的提取

模板的格式牵涉到整个匹配过程，所以模板格式的设计比词库和标注的设计要求更为严格。在得到人工标注的关键词序列之后，出于简化模板、增加模板匹配成功率的目的，系统根据标点符号将一句复杂的文本分成几个简单句，然后对每个简单句分别提取模板。为了必要的模板匹配，首先需要按顺序保存每个关键词的词性和类别。考虑到如果将一句评论性语句中的褒义词语换成贬义词语，那么整句文本的褒贬倾向也将改变，所以对于模板，褒贬倾向是不必保存的，进而程度也不必保存。因此，表 1 中的关键词序列提取出的模板为/d2/a1。目前系统的模板库中模板数目为 86 个。此外，系统还保存了模板在训练文本库中出现的频率。

3.2 模板的匹配

模板匹配是整个系统的核心算法，为了提高匹配的准确度和效率，本文提出了一个层次化模板的概念，将复杂的文本拆分为几个简单句，分别进行模板匹配，而最终文本的褒贬倾向值由各个简单句的褒贬倾向值相加得到，褒贬倾向值的计算下一章将详细介绍。此外，在模板匹配过程中，系统首先进行关键词类别的匹配，以缩小匹配范围，简单起见，本文称关键词类别序列为类别模板。具体的匹配算法流程如下：

针对一句输入文本，将它分为简单句，对于每一个简单句，提取出类别模板，如果提取出的类别模板为空，转 5)；

把简单句的类别模板同模板库中所有类别模板进行匹配，如果有一个或多个类别模板完全匹配，选择频度最高的作为匹配结果，转 4)。否则，转 3)；

再将简单句的类别模板同模板库中所有类别模板比较，对于模板库中每一个模板 M，如果 M 的类别模板包含简单句的类别模板，那么在匹配点的区间之内进行词性的匹配，如果词性匹配上且分类为 1，计算文本对应位置词语的褒贬倾向。最终选择匹配词性最多的模板作为匹配结果，如匹配词性的数目相同，选择频度大的模板匹配。如果找不到符合条件的 M，转 5)；

根据匹配的结果，计算简单句的褒贬倾向。转 1，计算下一个简单句的褒贬倾向，并累加；

匹配失败，转 1，计算下一个简单句的褒贬倾向，并累加。

表 3 为一个模板匹配过程的例子。在原始文本“今天电影很精彩”中提取简单模板 21，发现在库中有两个模板能够完全匹配，于是最终选择频度最高的模板进行匹配。

4　词语文本褒贬倾向的计算

在模板匹配过程中，如果遇到关键词词库之外类别为 1 的词语，系统需要计算它的褒贬倾向。而在模板匹配完成之后，系统还需要计算整句文本的褒贬倾向。

表 3 模板匹配举例

Table 3 An example of template matching

模板匹配步骤	模板匹配结果
原始文本	今天电影很精彩。
关键词匹配结果	今天/t　电影/n　很/d2#2　精彩/a1+2　。/w
类别模板	21
满足匹配条件的模板	/d2/a1 频度：5
	/i2/a1 频度：2
最终匹配的模板	/d2/a1

4.1 词语褒贬倾向的计算

词语褒贬倾向的计算方法为选取一定数量的褒义词和贬义词作为参考词汇(一方面按照经验选取一些经典的褒义词和贬义词，另一方面则到关键词词库里选取频度高的 1 类关键词)，计算该词与参考词汇的相似度，选出与褒义词相似度的最大值和与贬义词相似度的最大值，求差并与阈值做比较，大于阈值为褒义词，小于阈值为贬义词，简单起见，本系统中将阈值取为 0。

计算词汇相似度算法的设计本文参照了文献[6]并对下述问题进行了改进。

式 1 为原论文中词汇相似度的计算公式，该式的主要思想为主要部分的相似度值对于次要部分的相似度值起到制约作用。但是，最后对四部分结果加权求和时，可能存在四部分的某一部分在两个义项中都不存在的情况，按照式 1 计算会造成结果偏小，这是不合理的。因此本系统在计算前先判断每部分在两个义项中是不是都不存在，如果都不存在的话就去掉相应的β 值，重新按比例调整其他β 值使它们的和仍为 1。

$$Sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}Sim_j(S_1,S_2) \qquad \cdots\cdots (1)$$

4.2 文本褒贬倾向的计算

在模板匹配成功之后，需要根据一定的规则计算出整句文本的褒贬倾向。这个规则的设定需要在一定程度上体现出语法规则，否则将很容易导致计算出的整个语句的情感倾向错误。例如，程度副词既可能出现在其中心词的左侧，也可能出现在其中心词的右侧（“很好”，“好得很”）。本系统文本褒贬倾向计算规则设定如下：

1) 根据模板从文本中取出所有模板成分对应的词，去掉不相关的词，组成一个序列

2) 第一遍扫描序列，找到所有程度副词(类别为 2)，将其程度值乘到模板中离其最近的一个 1 类词的程度值上(考虑到副词可能位于其中心词的前面或者后面，所以

这里的“最近”是前后双向的查找，同时由于副词在前的情况比较多，所以前向查找的优先级高)。具体的处理是标注程度为 3 的因子为 1.5,程度为 2 的因子为 1，程度为 1 的因子为 0.5。

3) 第二遍扫描序列，找到所有否定词(类别为 3)，将其往后碰到的第一个 1 类词的褒贬性取反。

4) 第三遍扫描序列，以转折词为单位将序列分成几个小部分，对每个小部分累加其 1 类词的褒贬倾向值，然后按转折词类型的不同乘以转折词相应的权值(让步型如“虽然”，对应部分要减弱；转折型如“但是”，对应部分要加强)，最后各部分相加得到文本的褒贬倾向值。

表 4 中文本计算得到的褒贬倾向值为 2,即最终判定为褒性评论。

5 实验分析

为了检验系统的性能，本文对当前文本褒贬倾向判定系统进行了集内和集外测试，当前系统的训练集为 100 句电影《变形金刚》的评论，其中褒义评论 69 句，贬义评论 31 句。集内测试结果如表 3。

表 3 系统集内测试结果

Table 3 System close test result

文本类别	文本总数	判断正确	判断错误	正确率%
褒义文本	69	59	10	85.5
贬义文本	31	27	4	87.1

可以看到，系统对于集内测试，无论测试文本的褒贬性，正确率都比较高。这是因为对于集内测试，系统可以很好的进行关键词以及模板的匹配,进行可以准确的计算文本的褒贬倾向。这是由于系统首先在关键词匹配阶段从文本中提取出了对文本褒贬倾向有影响的词汇,接下来通过模板匹配来判断上一步提取的词汇集是否可以合理的完整的表达褒贬倾向，并且对于关键词词库之外的含有褒贬倾向的词汇，系统还设计了利用词汇相关度计算词汇褒贬倾向的方法,最终,系统按照一定的规则计算出整句文本的褒贬倾向。

除了集内测试之外，文本还对系统进行了集外测试，然而测试结果很不令人满意，重要原因是关键词词库过小。比如对于测试文本：

昙花一现又乏善可陈的戏份让他们尽数沦为过眼云烟。

由于关键词词库中没有“昙花一现”、“乏善可陈”而导致关键词和模板匹配失败，最终系统无法判断文本的情感。解决这类问题的主要方法应为扩充词库。

另外，当前系统使用的模板提取及匹配方式，难以处理语义复杂的文本。例如对于测试文本：

两个多小时的电影就是一次电脑特效与电影的联姻，之前没有任何电影能够企及这样的高度。

这是一句褒义评论，通过对“任何电影”做贬义评论，表达对主语“两个多小时的电影”的褒义评论。但当前系统不能分辨出后半句的评论对象并非主语,而导致最终判定倾向为贬义，出现错误。为了解决这个问题，就需要在模板中引入更多的语义信息。

6 结语

本文设计了一个基于关键词模板的文本褒贬倾向判定系统。该系统通过文本句法分析、关键词匹配、模板匹配、词汇褒贬倾向计算以及文本褒贬倾向。

进一步的工作是扩大关键词词库的规模,增加领域相关特性、在模板中引入更多的语义信息等，以提高文本褒贬倾向判定的精度。

References:

[1] Zhu Yan-lan, Min Jin, Zhou Ya-qian、Semantic Orientation Computing Based on HowNet、Journal of Chinese Information Processing，2006，20（1）：14-20.

[2] Peter D. Turney、Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews、the Association for Computational Linguistics 40th Anniversary Meeting，New Brunswick，N.J.，2002：417~424.

[3] Wen Tao，Zhu Qiao-ming，Lv Qiang、A Fast Algorithm for Chinese Word Segmentation、Computer Engineering，2004, 30（19）：119-120，182.

[4] Zhao Wei，Dai Xin-yu，YIN Cun-yan、A Method Combining Rule-based and Statistics-based Approaches for Chinese Word Segmentation、Application Research of Computers，2004，21（3）：23-25.

[5] Liu Qun、Chinese Lexical Analysis System、SWCL，2002.

[6] Liu Qun，Li Su-jian、Semantic Similarity Computing Based on HowNet，CLSW2，2002：59-76.

[7] HowNet. http://www.keenage.com. 2007.

附中文参考文献：(小 5 黑)

[1] 朱鄢岚，闵锦，周雅倩、基于 Hownet 的词汇语义倾向计算、中文信息学报，2006，20（1）：14-20.

[3] 温滔，朱巧明，吕强、一种快速汉语分词算法、计算机工程，2004，30（19）：119-120，182.

[4] 赵伟，戴新宇，尹存燕，陈家骏、一种规则与统计相结合的汉语分词方法、计算机应用研究，2004，21(3)：23-25.

[5] 刘群、汉语词法分析系统、第一届学生计算语言学研讨会，2002.

[6] 刘群，李素建、基于《知网》的词汇语义相似度计算、第三届汉语词汇语义学研讨会论文集，台北，2002：59-76.

[7] 知网，http://www.keenage.com. 2007.