文章编号：1003-0077(2012)05-0014-06

用户评论中的标签抽取以及排序

李丕绩[1]，马 军[1]，张冬梅[2]，韩晓晖[1]

(1. 山东大学 计算机科学与技术学院，山东 济南 250101
2. 山东建筑大学 计算机科学与技术学院，山东 济南 250101)

摘 要：对于一个实体(产品或者商户)，往往伴随着成千上万的用户评论。如何从这些冗杂的评论信息中抽取能够描述此实体的精华信息是研究的热点问题。该文提出了一种能够为每个实体抽取特征标签的方法，并且语义去重，保证标签在语义空间内相互独立。首先，对于每个实体的所有评论，进行中文分词、词性标注，并且做依存句法分析。然后，根据每个句子中的依存关系，抽取关键标签，构成此实体的标签库，并且对标签库进行显式语义去重。最后通过 K-Means 聚类以及 Latent Dirichlet Allocation(LDA)主题模型将每个标签映射到语义独立的主题空间，再根据每个标签相对该主题的置信度进行排序。通过以上步骤，可以为每个实体抽取语义独立的关键标签描述，实验中，该文通过对返回标签列表的准确性以及语义多样性进行了统计分析，验证了标签抽取方法的可行性和有效性。

关键词：意见挖掘；主题模型；语义独立；标签抽取；排序

中图分类号：TP391 **文献标识码**：A

Extraction and Ranking of Tags for User Opinions

LI Piji[1], MA Jun[1], ZHANG Dongmei[2], HAN Xiaohui[1]

(1. College of Computer Science and Technology, Shandong University, Jinan, Shandong 250101, China
2. College of Computer Science and Technology, Shandong Jianzhu University, Jinan, Shandong 250101, China)

Abstract: There are usually millions of comments for an entity (e. g. a shop or a product). How to extract the consice and useful information to describe the entity is a challenging issue. This paper proposes a method to extract tags without semantic redundancy. First, we perform the word segmentation, POS tagging and dependency parsing for all the comments. Then, we extract tags acroding to the dependency realtions, and reduce the semantically duplicate tags explicitly. Finally, we map all the tags to the independent semantic space via K-Means and Latent Dirichlet Allocation(LDA), and rank the tag list. according to the topic confidence. The results of the experiments show that our method could extract the tags accurately with semantic independency.

Key words: opinion mining; topic model; semantic independent; tag extraction; ranking

1 前言

近年来，互联网和电子商务的飞速发展不仅给企业的业务流程带来了巨大的变革，而且对消费者的行为模式也产生了深刻的影响。因此，网络上各种产品的评论数量也在飞速地增长。而且越来越多的证据表明，评论信息影响到消费者的购买决定。但是，随着时间的推移，产品的评论会越来越多，评论列表会变得很长或者分很多页。随着评论数量的积累，评论的质量也会参差不齐。这样，用户在浏览评论的时候就会花费很多的时间和精力，甚至看不到隐藏在网页最深处的而且可能是最有价值的评论信息。

收稿日期：2011-09-15 定稿日期：2012-03-21

基金项目：国家自然科学基金资助项目(60970047，61103151，61173068)；教育部博士点基金资助项目(20110131110028)

作者简介：李丕绩(1986—)，男，主要研究方向为信息检索；马军(1956—)，男，教授，博士生导师，主要研究方向为信息检索；张冬梅(1970—)，女，副教授，主要研究方向为信息检索。

那么，能否从这些成千上万的评论中抽取对这个产品有效准确的简短描述，能够让用户最快时间获得此产品的重要信息呢？

为了解决这个问题，本文提出了一种能够为每个实体①抽取特征标签的方法，并且语义去重，保证标签在语义空间内相互独立。首先，对于每个实体的所有评论，进行中文分词、词性标注，并且做依存句法分析。然后，根据每个句子中的依存关系，抽取关键标签，构成此实体的标签库，并且对标签库进行显式语义去重。在本文中，对关键标签的抽取主要关注对实体某些属性的实际描述词，例如，“味道不错”“价格实惠”等，显式去重是根据预先定义的同义词词典进行初步去重，例如，“口味儿”“味道”都看作“味道”。最后，通过 K-Means 聚类以及 Latent Dirichlet Allocation(LDA)[1] 主题模型将每个标签映射到语义独立的主题空间，从每个主题空间中抽取单个标签，再根据每个标签相对该主题的置信度和支持度进行排序。通过以上步骤，可以为每个实体抽取语义独立的 N 个关键标签描述。实验中，本文通过对返回标签列表的准确性以及语义多样性进行了统计分析，验证了标签抽取方法的可行性和有效性，并且有一定的实际应用价值。

2　相关工作

产品评论挖掘的一个主要任务是需要了解用户对产品的哪些功能、部件和性能进行了评价，因此需要从产品评论中提取出用户评价的对象——产品特征。用户在产品评论中对特征的描述，可能是厂家根本没有考虑到的一些特征，因此挖掘出产品评论中所提及的特征，了解用户对这类产品最关心的功能和性能是具有重要意义的。

产品特征的提取分为人工定义和自动提取两类。在人工定义方面，Kobayashi、Inui 和 Matsumoto[2] 以人工定义方式构建了针对汽车的产品特征，共有 287 个产品特征，每一个特征使用一个三元组进行表示(＜Attribute，Subject，Value＞)，其中 Subject 表示产品，Attribute 表示产品的特征，Va-lue 表示对这个特征的观点；姚天昉[3-4] 利用本体建立了汽车的产品特征，该系统可在电子公告板、门户网站的各大论坛上挖掘并且概括意见持有者对各种汽车品牌的不同性能指标的评论和意见，并且判断这些意见的褒贬性以及强度；Li Zhuang[5] 针对电影评论人工定义了电影的产品特征，将电影的产品特征分为两类：电影的元素(screen play，vision effect) 与和电影相关的人员 (director，screenwriter，actor)。

自动提取产品特征的方法，需要使用词性标注、句法分析和文本模式等自然语言处理技术对产品评论中的语句进行分析。自动发现产品特征，由于不需要大量的标注语料库作为训练集，因此具有较好的通用性，并且可以适用于各种产品，可以比较容易地移植到不同产品上，但它最大的缺点就是准确率比较低。Hu 和 Liu[6] 先对评论语料进行词性标注，然后把每个句子中的名词和名词短语提取出来，利用关联规则挖掘方法从评论语料中取出满足最小支持度的名词或名词短语生成 transaction file，再使用 CBA(Classification Based on Associations)[7] 从 transaction file 中挖掘出频繁项，把频繁项作为产品特征候选集，由于关联规则产生的频繁项不是全都是有用的或真正的特征词，需要进行进一步的筛选，首先去掉了三个词以上的名词短语，然后对候选特征集中的候选特征进行修剪，通过“紧凑修剪”和“冗余词修剪”移除那些很大可能不是产品特征词的名词短语。

Popescu[8] 把评论挖掘分成四个主要子任务：(1)识别产品特征；(2)识别产品特征对应的观点词；(3)判断观点词的极性；(4)根据观点的强度排序。他们在 KnowItAll[9] 网络信息抽取系统基础之上建立了一个无监督的信息挖掘系统 OPINE。在产品特征识别方面，Popescu 建立的 OPINE 系统将产品特征分成显式特征和隐式特征，其中显式特征又分为五类，分别为“properties、parts、features of productparts、related concepts、parts and properties of related concepts”。用 OPINE 来挖掘产品特征的准确率比 Hu [12] 挖掘结果高出了近 22%，而召回率仅下降了 3%。

在本文中，由于目标是要用三至五个关键词对商户进行特征的描述，所以本文在基于句法分析的基础上，提出了基于 K-Means[10] 和基于 Latent Dirichlet Allocation (LDA)[1] 的两种关键词抽取方法，经过试验，效果十分理想。

3　方法

假设已经得到了某个实体的所有评论信息，算

① 实体：本文的实体指的是网络中存在的产品、商户或者店铺等。

法由此开始，可以分为四个子模块：词法句法分析、候选标签的挖掘抽取、标签去重和语义独立、代表标签的选取以及排序。

3.1 词法句法分析

为了能够后续步骤中的候选标签抽取，需要对每个评论的每个句子进行词法和句法分析。将句子分词，并且进行词性标注，而且需要将词与词之间的修饰关系描述出来。

本文中我们使用依存句法分析来确定词之间的关系。例如，评论："服务员很漂亮。饭菜很好吃，味道不错。"进行句法分析之后的结果如图1所示。

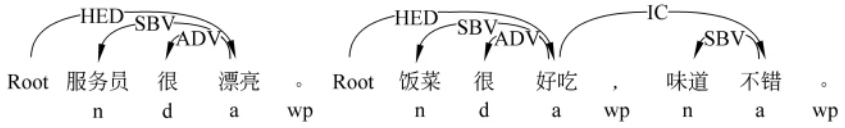

图1 依存句法分析例子

依存句法分析之后，就可以将有用的三元组<主题词，ADVs，修饰词>提取出来，作为描述此实体的一个标签。

3.2 候选标签抽取

将每个实体的所有评论的句子进行了词法句法分析之后，就可以进行候选标签的抽取，同时，根据初始词典进行显式语义去重。本文主要分析<主题词(n)，(ADVs)，修饰语(a)>三元组，也就是主要是考虑用形容词来修饰名词的标签信息。例如，"服务员很漂亮""价格很便宜"等。

通过句法分析得到了依存关系，对每个商户 s，就可以挖掘其候选标签集合 O，

$$O = \{O_1, O_2, \ldots, O_n\}$$
$$O_i = \{o_1, o_2, \ldots, o_m\} \quad (1)$$

其中 n 为商户 s 评论的数目，O_i 为第 i 个评论所产生的标签候选。

为了在后续的语义去重中更加准确，这里对于每个新的标签 o_i，都要进行显式去重。意思就是将重复的 o_i 过滤，并且根据初始的词典将显式语义进行合并。例如，"口味儿""味道"的主题标签都用"味道"的主题标签代替。

3.3 标签语义去重

对于商户 s 和其所有候选标签集合 O，怎么进行标签语义去重和保证相互独立呢？这是最重要的算法模块，本文提出了两种解决方法，并比较了两种方法的结果：(1)K-Means 主题聚类；(2)LDA 主题分析。

3.3.1 基于 K-Means 主题聚类

假设主题数目是 K。由于用户评论的相对稀疏性，所以 K 一般选择较小的值，例如，10。将 O 中的每个候选标签 o_i 看作一个单独的文档，然后对这 $|O|$ 个文档进行 K 聚类。

在聚类过程中，一个关键的问题是标签 o_i 和标签 o_j 的相似度计算，以及标签 o_i 和聚类 c_i 的距离，由于本文把每个标签看作一个文档，所以如果使用 tf-idf 等向量空间模型表示将会十分稀疏。为了避免这种稀疏性，一种方法是文档表示用基于字的 1-gram 表示成向量空间模型，这种方式在基于 LDA 的主题建模用到了。另一种方法把一个标签 o_i 看作一个字符串处理。

既然是看作字符串，那么计算距离的时候本文采用 Levenshtein Distance[11] 来衡量字符串之间的距离。为了更准确地达到主题聚类和消除语义重复的作用，只用标签的主题词 $t(o_i)$ 代替标签计算距离 LD。

$$LD(o_i, o_j) = LD[t(o_i), t(o_j)] \quad (2)$$

用这种方法，就无法计算聚类的中心。所以计算一个文档 o_i 到聚类 c_i 的距离，就近似用到这个聚类中所有文档的距离的均值才衡量。

$$Distance(o_i, c_i) = \frac{1}{|c_i|}\left[\sum_{j=1}^{|c_i|} LD(o_i, o_j)\right] \quad (3)$$

试验中迭代 20 次之内便会收敛。这样每一个聚类就可以看作描述某一个主题的标签的集合。

3.3.2 基于 LDA 主题模型

LDA(Latent Dirichlet Allocation)[1] 是一个多层的产生式概率模型。它有词、主题、文档三层结构。LDA 将每个文档表示为一个主题混合，而每个主题是固定词表上的一个多项式分布。本文中，由于把每个标签看作每个文档，为了解决稀疏性，用基于字的 1-gram 建立向量空间模型。经过估计和推断之后，对于每个商户，有了所有的文档—主题分布"θ"，这样就将所有的标签映射到不同的主题 z 上，

再从主题 z 中选择有代表性的标签作为代表输出即可。

3.4　代表标签选择及排序

本文以相互比较和互补的形式提供了三种选择代表标签的策略：K-Means Topic Clustering(KM-TC)、LDA Max Topic(LDA-MT)和 LDA Topic Clustering(LDA-TC)。下面详细描述每种方法。

3.4.1　K-Means Topic Clustering(KM-TC)

第一步，为每一个聚类 c 选择一个代表性的标签 o；

第二步，对选择出的 K 个标签 o 进行排序输出；

这样，就需要每一个标签有一个分数值 $Score(o)$，聚类 c_i 中的标签的分数，

$$Score(o_i,c_i)=e^{-\lambda[Distance(o_i,c_i)]}+w_i\frac{|c_i|}{|O|} \quad (4)$$

其中 $|O|$ 是该实体的标签总数。$\frac{|c_i|}{|O|}$ 表示如果这个聚类中的标签越多，那么这个聚类越重要，对 $score$ 有贡献。其实 $e^{-\lambda[Distance(o_i,c_i)]}$ 可以理解为置信度，$\frac{|c_i|}{|O|}$ 可以理解为支持度。

每个聚类 c 选择出一个 $score$ 最大的标签 o 作为代表，那么就会得到 K 个标签。

$$O_{best}=\{o_1,o_2,\ldots,o_{|K|}\} \quad (5)$$

并且按照 $score$ 做了排序，$score(o_1)\geqslant socre(o_2)\geqslant\ldots\geqslant score(o_{|K|})$。

3.4.2　LDA Max Topic(LDA-MT)

LDA 主题模型将每个候选标签都映射到主题分布空间中，为了将某个候选标签 o_d 赋予某个主题 z_i，采取了一种贪婪的策略：

$$SignTopic(o_d)=\underset{z_i\in Z}{\operatorname{argmax}}P(z_i\mid o_d) \quad (6)$$

实际上是将概率最大主题 z_i 作为标签 o_d 所属于的主题。

和第一种方法类似，也需要有个打分函数：

$$Score(o_i,z_i)=P(z_i\mid o_i)+\frac{|z_i|}{|O|} \quad (7)$$

其中 $P(z_i|o_i)$ 可以看作置信度，$\frac{|z_i|}{|O|}$ 可以看作支持度。这一过程其实也类似一种聚类，只是有点贪婪的思想。这样，对标签指派完主题之后，从每个主题中选出一个标签，然后再排序就得到了结果。

3.4.3　LDA Topic Clustering(LDA-TC)

主题模型最终将文档映射到主题分布 $P(z_i|o_i)$，可以将分布 $P(z_i|o_i)$ 看作一个主题维 K 的描述此标签信息的特征向量，这也可以看作一个由词典维映射到主题维的降维过程，那么有了特征描述，可以继续进行聚类，将相同潜在主题的标签映射到相同的主题上。

聚类中距离的计算我们采用欧氏距离：

$$Distance(o_i,o_j)=ED[P(z_i|o_i),P(z_j|o_j)] \quad (8)$$

代表标签选择和排序的过程也和 KM-TC 方法类似。

4　实验

4.1　数据集以及实现

为了验证方法的有效性，我们从大众点评网(www.dianping.com)上抓取了近 1 000 个商户的信息，所有商户的评论近 130 000 条。

在实现中，词法句法分析我们使用的哈尔滨工业大学的开源语言技术平台(Language Technology Platform，LTP)[12]①。LDA 主题模型我们使用的开源的 Java 版的 JGibbLDA②。

4.2　评测标准

本文的目的是要为每个商户抽取语义独立的标签进行准确描述，所以主要从准确性、语义多样性以及标签质量三个方面对结果进行了评测。准确性采用信息检索中常用到 P@n 和 MAP 作为评价指标。

4.2.1　D@n(语义多样性@n)

因为标签抽取中保证语义独立性非常重要，所以本文用语义多样性来度量，可以如下进行计算：

$$D@n=\frac{NR_{rel_n}}{N_{rel_n}} \quad (9)$$

其中 NR_{rel_n} 表示前 n 个准确的标签中语义独立的标签的数量，N_{rel_n} 表示前 n 个标签中准确描述的数量。

4.2.2　AQ(平均质量)

对于质量的度量，一般意义来说，标签的长度越长，包含的字数越多，可能隐含的语义和信息就更加完整，质量较高，所以可以用标签的平均长度进行度量。

$$AQ=\frac{1}{K}\sum_{i=1}^{K}\frac{length(o_i)}{\mathrm{Max}Length(O)} \quad (10)$$

① http://ir.hit.edu.cn/ltp/

② http://jgibblda.sourceforge.net/

其中 $length(o_i)$ 表示标签 o_i 的长度，用 $\mathrm{Max}Length(O)$ 进行归一化。

4.3 实验结果

试验中我们对每个商户都生成了若干个语义独立的标签进行描述，由于最终的结果需要专家进行评测，所以我们随机的选择了 50 个商户分别求其各个衡量指标，最后求均值得到最后的评测结果。经过试验，发现当 LDA 主题数目以及聚类数目 $K=15$ 左右，效果较好。所以在试验中我们令 $K=15$，文章不再对 K 进行讨论。

如图 2 所示，方法 KM-TC 和 LDA-MT 在描述的准确率方面相差不大，因为都是相当于对于不同的主题词以及不同的主题进行了处理，并且选择代表性的标签，所以错误率小；但是方法 LDA-TC 的结果却不甚理想，理论上该方法应该会胜于前两种方法。经过分析发现由于在 LDA-TC 中 K-Means 聚类时候使用了欧氏距离，没有处理数据的稀疏性带来的巨大影响，并且没有对标签的长度进行加权，所以结果会逊色。

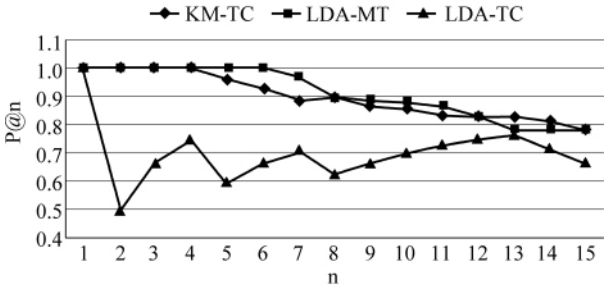

图 2 标签准确率

图 3 的左图是结果的 MAP 值，因为 MAP 与 P@n 有一定的关系，所以产生的结果也在预想之中，方法 KM-TC 和 LDA-MT 要优于 LDA-TC 方法。但是考虑到标签的质量问题，如图 3 的右图所示，发现第二种方法 LDA-MT 的质量要明显高于其他两种方法，而 LDA-TC 方法的质量变得非常差。分析其原因，由于 LDA-MT 对于每个文档是选择了其概率最大的主题作为其聚类的主题，所以如果该标签长度较长，含有的信息较多，那么其对于某个主题的贡献就越大，所以根据打分函数，得分也会较高。而方法 LDA-TC 的标签多为字数最少的标签，所以结果会很差。

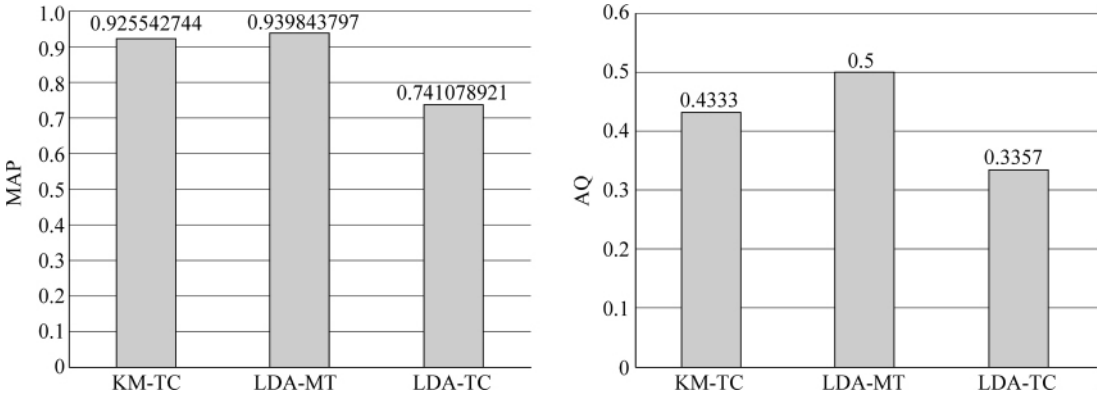

图 3 左：平均准确率 MAP；右：平均质量 AQ

对于这种情况的解决办法可以对标签的长度进行加权。一般来说，标签越长，含有的描述信息就越具体，越全面。

本文一个重要的问题是语义独立的问题，本文用语义多样性的指标 D@n 进行衡量，结果如图 4 所示。LDA-MT 因为标签的长度较长，含有的信息较全面，那么标签之间独立的可能性就比较大，所以其语义多样性会高于其他方法。但是方法 KM-TC 方法的语义多样性下降较大。分析其原因是因为在进行 K-Means 聚类的时候对标签使用了 LD 编辑距离，那么这其实是一种显式语义独立方法，但是不能揭示隐含的语义信息，例如，“苹果”和“电脑”的语义距离的计算。解决方法可以使用其他的语料库或者 wordnet 等来代替 LD 计算标签之间的语义距离。

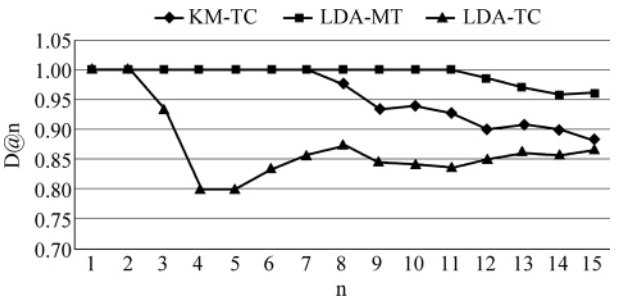

图 4 语义多样性 D@n

表 1 中我们给出了几个商户的三种方法的标签结果，由于版面原因我们只给出两个，另外我们还开发了一个标签抽取系统，以验证在实际应用中的可行性。

表 1　最终结果举例

商户名称和 ID	推荐标签结果（只举例前 5 个）
查餐厅（思南路店） （ID：3259888）	KM-TC：菜量很多 玫瑰不嫩 口感很好 气氛很好 鸡肉很嫩 LDA-MT：服务员很生硬 外国人特别多 觉得性价比高 虾仁很新鲜 奶茶很特别 LDA-TC：人多 餐厅和不同 味道好 美味不错 油不嫩
宜家家居（漕溪路店） （ID：1862461）	KM-TC：场地很好 周末人很多 态度很好 质量很好 利用率很好 LDA-MT：人流量越来越大 东西蛮精致 标准品完全合适 热狗很便宜 厨房间太小 LDA-TC：人多 东西蛮 价格贵 性价高 蛋糕不错
汉泰东南亚风味餐厅（徐汇店） （ID：2960611）	KM-TC：总体很好 分量都少 味道不错 态度很好 豆一般 LDA-MT：态度超级好 价格中便宜 服务员不多 粉丝很不错 豆一般 LDA-TC：芒果超级好吃 态度很好 肉很嫩 味道不错 豆一般
港丽餐厅（大悦城店） （ID：2384860）	KM-TC：吃饭人很多 美味很多 胃口不好 环境不好 肉质很好 LDA-MT：性价比较低 西餐厅野菌不同 口感真是神奇 丝瓜还不错 服务员很热情 LDA-TC：量大 价钱贵 菜好吃 肉嫩 味不够

5　总结

本文提出了为产品或者商户生成语义相关描述标签的方法。通过 K-Means 聚类以及 Latent Dirichlet Allocation(LDA)主题模型将每个标签映射到语义独立的主题空间，再根据每个标签相对该主题的置信度进行排序。实验可以看出，该种方法能够解决一定的实际问题。

在未来的工作中，需要继续考虑在聚类过程中语义距离的度量问题。此外，还需要考虑在时间维度上的动态主题迁移问题。

另外，对于目前社交网络中的用户对于产品或者商户的评价，可以将用户的信息以及其社交信息考虑到模型当中，进一步提高结果的准确性。

参考文献

[1] Blei D. M., A. Y. Ng, M. I. Jordan. Latent dirichlet allocation [J]. The Journal of Machine Learning Research, 2003. 3: 993-1022.

[2] Kobayashi N., K. Inui, Y. Matsumoto, et al. Collecting evaluative expressions for opinion extraction [C]//Proceedings of Natural Language Processing-IJCNLP 2004, 2005: 596-605.

[3] 姚天昉，聂青阳，李建超，等. 一个用于汉语汽车评论的意见挖掘系统[C]//中文信息处理前沿进展——中国中文信息学会二十五周年学术会议，北京：清华大学出版社，2006：260-281.

[4] 姚天昉，程希文，徐飞玉，等. 文本意见挖掘综述[J]. 中文信息学报，2008，22(3)：71-80.

[5] Zhuang L., F. Jing, X. Y. Zhu, et al. Movie review mining and summarization [C]//Proceedings of the 15th ACM International Conference on Information and Knowledge Management 2006: 43-50.

[6] Hu, M., B. Liu. Mining opinion features in customer reviews[C]//Proceedings of 19th National Conference on Artificial Intelligence: Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. 2004: 755-760.

[7] Ma B. L. W. H. Y. Integrating classification and association rule mining [C]//Proceedings of In Knowledge Discovery and Data Mining, 1998.

[8] Popescu A. M., O. Etzioni. Extracting product features and opinions from reviews[C]//Proceedings of HLT-Demo '05 HLT/EMNLP on Interactive Demonstrations Association for Computational Linguistics. 2005: 339-346.

[9] Etzioni O., M. Cafarella, D. Downey, et al. Unsupervised named-entity extraction from the web: An experimental study[C]//Proceedings of Artificial Intelligence, 2005: 165(1): 91-134.

[10] MacQueen J. Some methods for classification and analysis of multivariate observations [C]//Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. California, USA, 1967: 14.

[11] Levenshtein Distance [OL]. http://en.wikipedia.org/wiki/Levenshtein_distance.

（下转第 45 页）

该检索系统目前只是集中解决了单问句的相似度匹配问题。对于一个含有多个问句，对问句进行大量篇幅的说明，所包含的信息量很大的提问，本检索系统没有得到很好的应用，这有待于对检索框架的进一步改进和完善，这也是今后继续研究的方向。

参考文献

[1] A. Berger, R. Caruana, D. Cohn, et al. Bridging the Lexical Chasm: Statistical Approaches to Answer-Finding[C]//Proceedings of SIGIR, New York, NY, USA, 2000: 192-199.

[2] Song Wanpeng, Feng Min, Gu Naijie, et al. Question Similarity Calculation for FAQ Answering [C]// Proceedings of SKG, 2007: 298-301.

[3] D. Molla, J. Vicedo. Question answering in restricted domains: An overview[J]. Computational Linguistics, 2007, 33(1):41-61.

[4] J. Jeon, W. B. Croft, J. H. Lee, et al. A framework to predict the quality of answers with non-textual features[C]//Proceedings of SIGIR, Seattle, USA, 2006: 228-235.

[5] M. Blooma, A. Chua, D. Goh. A predictive framework for retrieving the best answer [C]// Proceedings of SAC, Brazil, 2008: 1107-1111.

[6] Cao Xin, Cong Gao, Cui Bin, et al. A Generalized Framework of Exploring Category Information for Question Retrieval in Community Question Answer Archives[C]//Proceedings of WWW, Raleigh, New York, NY, USA. 2010: 201-210.

[7] J. Ko, L. Si, E. Nyberg. A probabilistic framework for answer selection in question answering [C]// Proceedings of NAACL/HLT, Rochester, NY, 2007: 524-531.

[8] Wang Xinjing, Tu Xudong, et al. Ranking community answers by modeling question-answer relationships via analogical reasoning[C]//Proceedings of SIGIR, New York, NY, USA. 2009: 179-186.

[9] P. Jurczyk, E. Agichtein. Discovering authorities in question answer communities by using link analysis [C]// Proceedings of CIKM, New York, NY, USA, 2007: 919-922.

[10] Shen Jie, Shen Wen, Fan Xin. Recommending Experts in Q&A Communities by Weighted HITS Algorithm[C]//Proceedings of IFITA, 2009: 151-154.

[11] J. Zhang, M. Ackerman, L. Adamic. Expertise networks in online communities: Structure and algorithms[C]//Proceedings of WWW, New York, NY, USA, 2007: 221-230.

[12] Liu Yandong, Bian Jiang, E. Agichtein. Predicting Information Seeker Satisfaction in Community Question Answering [C]//Proceedings of SIGIR, New York, NY, USA. 2008: 483-490.

[13] M. Blei, A. Ng, M. Jordan. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, b: 993-1022.

[14] T. L. Griffiths, M. Steyvers. Finding scientific topics[C]//Proceeding of the National Academy of Sciences. 2004: 5228-5235.

（上接第 19 页）

[12] Che W., Z. Li, T. Liu. Ltp: A chinese language technology platform [C]//Proceedings of Coling 2010, Demonstrations: Association for Computational Linguistics. 2010: 13-16.