基于电子商务用户行为的同义词识别*

张书娟，董喜双，关 毅
哈尔滨工业大学，哈尔滨 150001
E-mail: zhangsj777@gmail.com; dongxishuang@gmail.com; guanyi@hit.edu.cn

摘 要：本文研究了电子商务领域同义词的自动识别问题。针对该领域新词多、错别字多、近义词多的用词特点，提出基于用户行为的同义词识别方法。首先通过并列关系符号切分商品标题和基于 SimRank 思想聚集查询两种方法获取候选集合，进而获取两词的字面特征以及标题、查询、点击等用户行为特征，然后借助 Gradient Boosting Decision Tree（GBDT）模型判断是否同义。实验表明同义词识别准确率达到了 54.46%，高于 SVM 近 4 个百分点。

关键词：同义词识别；用户行为；SimRank；Gradient Boosting Decision Tree

The Synonym Recognition Based on User Behaviors in E-commerce

Zhang Shu-juan, Dong Xi-shuang, Guan Yi
Harbin Institute of Technology, Harbin 150001
E-mail: zhangsj777@gmail.com; dongxishuang@gmail.com; guanyi@hit.edu.cn

Abstract: This paper focuses on synonym recognition in e-commerce. Considering there are more new words, typos, and near-synonyms in e-commerce domain, we present a method to recognize synonyms based on user behaviors. Firstly, candidate sets are retrieved by analyzing the titles and their corresponding queries based on SimRank theory, and then, features including literal feature, title feature, query feature, click feature are extracted. Finally, Gradient Boosting Decision Tree(GBDT) model is adopted to determine whether candidate synonyms are real synonyms or not. The experimental result shows that GBDT model is more suitable for this task with precision 54.46%, which is nearly 4 percent higher than that of SVM.

Keywords: synonym recognition; user behaviors; SimRank; Gradient Boosting Decision Tree

1 引言

随着互联网的发展，电子商务逐步发展起来。对于电子商务网站的站内搜索引擎而言，应该在准确理解买方意图的基础上，尽可能多的检索出相关商品，所以需要对查询进行扩展。要准确扩展查询，同义词表是必须的也是最基础的资源。目前国内还不具备电子商务领域的同义词典，而手工构建又费时费力，所以需要采用同义词自动识别的方法。

牛津字典对同义词的定义为：在用同一种语言表达的意义相同或者相近的两个词或者短语[1]。而电子商务中则要求意义完全相同，定义为对同一事物或者概念的不同表达，即在商品检索和商品描述中可以互相替换的词[2]。电子商务领域中同义词主要有六类：（1）中英文名称，如：耐克-Nike。（2）学名与俗名，如：圣女果-小番茄。（3）全称与简称，如：美特斯邦威-美邦。（4）新称与旧称，如：自行车-脚踏车。（5）常用错别字引起的同义，如：瑜伽-瑜珈。（6）传统同义词，如：储物柜-收纳柜。

电子商务领域同义词的特殊性，使现有自动识别方法的效果大打折扣。一方面现有资源中给出的同义词不一定满足此领域中的定义，例如在《同义词词林》中木耳与黑木耳为同义词，其实两者是上下位关系，木耳除了包括黑木耳之外还有秋木耳、白木耳等。另一方面网络用词中新词多、错别字多，大量的词现有资源还未收录。因此，需要根据电子商务领域数据特点寻找新的同义词的识别方法。

* 本项研究受到国家自然科学基金项目支持，项目批准号：60975077，60736044。

本文在研究电子商务中卖方用户和买方用户行为特点的基础之上，提出了基于用户行为的同义词自动识别方法。首先根据用户行为特点获取候选集，进而提取两词的字面特征以及标题、查询、点击等用户行为特征，然后借助 GBDT 模型判断是否同义。结构组织如下：第二部分介绍国内外相关研究；第三部分介绍基于用户行为的同义词识别方法；第四部分分析实验结果；第五部分做总结和展望。

2 相关研究

目前国内外对同义词自动识别的研究，根据所使用资源可分为以下五类：

2.1 基于字面相似度

这种方法是根据两个词汇中相同字的个数来计算相似度。文献[3]根据词汇之间字面相似度将待归类词与被匹配词之间的聚类关系分为正确、不确定和无法判断三个级别，然后依赖专家对后两种情况进行判定，形成一种人机结合的同义词识别方法。文献[4]对上述方法进行了改进，根据汉语构词特点，引入重心后移，对词语中的每个语素根据其对主题表达的贡献进行加权处理，提高了准确率。

2.2 基于语义词典

这种方法借助现有语义词典或者自己构建语义词典来计算词汇相似度。文献[5]用自己建立的词素语义词典对待识别的词切分成多个词素，以计算两词汇相似度。文献[6]将《同义词词林》语义分类编码体系构成一棵树，通过计算树中结点距离得到词汇之间的相似度。文献[7]利用《知网》，在《知网》中每个词的语义由多个义原组成，将所有义原根据上下位关系构成一个树状层次体系，通过计算路径距离得到相似度，将两个词各自义原中相似度最大的作为两词的相似度。

2.3 基于词典释义

根据词典中词汇之间的相互注释关系，构造关系图，字典中的每个词都是图中的一个结点，词到它的每个注释都有一条边。将同义词的识别问题转化为互联网中超链接分析问题。文献[8]用 HITS 算法分析关系图，得到词汇之间相似度。文献[9]在 PageRank 算法的基础上提出 ArcRank 算法来计算词汇之间相似度。

2.4 基于大规模语料库

这种方法将词汇的上下文表示成空间向量。文献[10]将向量的余弦相似度作为两词的语义相似度。文献[11]在此基础上引入部分句法分析，只处理名词，在语料库中此名词的所有修饰词作为上下文，用 Jaccard 相似度来度量语义相似度。基于语料库方法所识别的同义词受语料库所属领域局限，且有数据稀疏的问题。

2.5 基于搜索引擎

这种方法借助搜索引擎的检索结果来统计词汇的出现次数，从一定程度上解决了统计的数据稀疏问题。文献[12]提出 PIM-IR 算法，通过计算互信息得到两词相似度。文献[13]对文献[12]的方法进行改进，提出了 LC_IR 算法，要求两词必须完全相邻，并且过滤搭配和修饰等噪声，提高了准确率。文献[14]则用 Dice 测度度量两词的相关性。

电子商务领域同义词与传统同义词定义的差异和新词较多的特点使得现有同义词自动识别方法的效果大打折扣。因此，本文在充分研究电子商务领域数据的基础上，根据用户行为特点获取

候选集合，然后提取字面相似度、共现信息、上下文、用户行为等方面的特征，运用机器学习方法对候选集合中的词对进行判定。

3 基于电子商务用户行为的同义词识别

电子商务中用户行为包括卖方用户行为和买方用户行为。本文主要研究卖方用户行为中的标题编辑行为，包括用词特点、词与词之间的分割方式等方面和买方用户行为中的查询和点击行为，包括查询中词的个数、词与词之间的分割方式、所点击的商品标题等方面。根据卖方行为特点从商品标题中获取候选集，并根据买方行为特点从查询集合中获取候选集，抽取部分候选进行标注，然后提取字面特征和标题、查询、点击等用户行为特征，最后训练 GBDT 模型以判定所有候选同义词对。

3.1 候选集合获取

3.1.1 并列关系符号切分商品标题

研究发现卖方在编辑商品标题时，除了写入商品常用名称之外，还会将该商品的别称、简称、全称、俗语、常用错别字等扩展写入标题之中，以使更多的买方检索到。并且标题多用空格、'/'、'\' 等表示并列关系的符号（称之为并列关系符号）分割表义相同或相近的词。研究某电子商务网站 3 百万商品标题，发现 72.4%的标题使用并列关系符号，因此我们用并列关系符号切分标题得到候选同义词对。

例如：对于商品标题“正品促销拉杆包/拉杆箱/旅行包/拉杆旅行包/旅行箱情侣搭配”，用并列关系符号切分得到拉杆包、拉杆箱、旅行包、拉杆旅行包、旅行箱五个词，两两组合行成候选词对。

3.1.2 基于 SimRank 思想聚合查询

SimRank 由 GJeh 和 J.Widom 于 2002 年提出，基本思想是关联到相似事物的两个事物相似[15] [16]。基于这一思想我们认为，点击到同一商品的所有查询相似。将点击到同一标题的所有查询聚合成查询集合，并从中获取候选同义词对。

研究某电子商务网站七天的查询日志（共 2 千万查询）发现关键词个数为 1 或者 2 的查询占总查询的 73.2%，而用空格分隔的查询占总查询的 89.4%。也就是说，大部分买方搜索商品时，仅使用简短的 1~2 个词汇进行搜索，且习惯于用空格自然分割查询。所以我们在查询集合内，用空格切分每个查询，去掉相同词段，剩余词段两两组合为候选同义词对。

例如：title^A 特价新款拉杆箱旅行箱行李箱密码箱托运箱 24 寸^A50012019

query^A 旅行箱^A1

query^A 拉杆箱^A1

query^A 箱^A3

query^A 行李箱^A1

标题数据格式：标记^A 标题^A 类目；查询数据格式：标记^A 查询^A 频率；^A 为分隔符。

旅行箱、拉杆箱、箱、行李箱这四个词两两组合行成候选词对。

3.2 特征提取

对于机器学习的分类方法而言，最重要的是选择一系列能够区分各类别的特征。由上文示例可见候选集合中大多是词义相近的词对，所以仅根据简单特征很难区分两词是否同义。因此本文在考虑简单字面特征的基础之上，着重选择用户行为相关的特征。经过实验，选择的特征主要包括以下四类：

（1）字面特征

同义词是指向同一事物的两个不同的词语，故常常含有共同的语素，例如，连身裤和连体裤，跑步鞋和跑鞋，因此考虑其字面相似度。网络用语常出现错别字，如运动品牌“阿迪达斯”一词，很多人由于输入错误使用“啊迪达斯”，当很多人都习惯于如此使用时，我们就不可以忽略这个问题，因此考虑两词的读音相似度。

$$Sim\_char_{min} = \frac{same(w_1, w_2)}{min(|w_1|, |w_2|)} \tag{1}$$

$$Sim\_char_{max} = \frac{same(w_1, w_2)}{max(|w_1|, |w_2|)} \tag{2}$$

$$Sim_{dis} = 1 - \frac{minDis(S_{w_1}, S_{w_2})}{max(|S_{w_1}|, |S_{w_2}|)} \tag{3}$$

其中，Sim_char_{min} 代表对较短词长的字面相似度，Sim_char_{max} 代表对较长词长的字面相似度，Sim_{dis} 代表读音相似度，$same(w_1, w_2)$ 代表在词 w_1, w_2 中相同字的个数，$|w_1|$ 代表词长，S_{w_i} 代表 w_i 的读音，$minDis(S_{w_1}, S_{w_2})$ 代表最小编辑距离。

（2）标题特征

如果两个词同义，根据卖方书写标题的习惯，应该大量出现在同一标题中，且两者的前后顺序应该是随机的。因此计算在所有商品标题中，两词共现比例，互信息和互换比例。互换比例用一个词总出现在另一个词前的概率来度量，这个特征可以很好的区分将修饰关系和同义关系。

$$Co\_ratio_{title} = \frac{2*C(w_1w_2)}{C(w_1) + C(w_2)} \tag{4}$$

$$MI_{title} = \log \frac{p(w_1w_2)}{p(w_1)p(w_2)} \tag{5}$$

$$Front\_ratio_{title} = \frac{C_f(w_1w_2)}{C_f(w_1w_2) + C_b(w_1w_2)} \tag{6}$$

以上各量都是对商品标题而言的，Co_ratio_{title} 表示两词共现比例，$C(w_1)$ 表示包含词 w_1 的数目，$C(w_1w_2)$ 表示同时包含词 w_1 和 w_2 的数目。MI_{title} 表示两词共现互信息，$p(w_1w_2)$ 表示两词共现频率，$p(w_1)$ 表示词 w_1 的频率。$Front\_ratio_{title}$ 表示两词互换比例，$C_f(w_1w_2)$ 表示 w_1 在 w_2 前面的数目，$C_b(w_1w_2)$ 表示 w_1 在 w_2 后面的数目。

（3）查询特征

同样考虑在查询中两词的共现比例，互信息和互换比例（计算公式同 title）。除此之外，还考虑每个词的上下文信息，即取这个词在查询中的前一个词和后一个词作为上下文，计算两词上下文的余弦相似度。

$$Sim_{cos} = \frac{V(w_1)*V(w_2)}{|V(w_1)|*|V(w_2)|} \tag{7}$$

Sim_{cos} 表示两词上下文向量的余弦相似度，$V(w_1)$ 表示 w_1 的上下文向量，$|V(w_1)|$ 表示 w_1 的上下文中词个数。

（4）点击特征

如果 w_1 出现在查询中，但没有出现在点击标题中，而 w_2 却出现在点击标题中，这种情况下两词很可能同义。因此需要考虑一个词出现在查询中，而另一个词出现在点击标题中这种共现的比例，互信息和互换比例。同时也需要考虑两个词都出现在标题中时，查询中只出现词 w_1 与只出现词 w_2 的比例。

$$Co\_ratio_{click} = \frac{C(w_{1t}w_{2q}) + C(w_{2t}w_{1q})}{C(w_{1t}w_{2q}) + C(w_{2t}w_{1q}) + C(w_{1t}w_{1q}) + C(w_{2t}w_{2q})} \tag{8}$$

$$MI_{click} = \log\frac{p(w_{1t}w_{2q}) + p(w_{2t}w_{1q})}{p(w_1)p(w_2)} \tag{9}$$

$$Front\_ratio_{click} = \frac{C(w_{1t}w_{2q})}{C(w_{1t}w_{2q}) + C(w_{2t}w_{1q})} \tag{10}$$

$$Query\_ratio_{click} = \frac{C_{q:}(w_1)}{C_q(w_1) + C_q(w_2)} \tag{11}$$

以上各量都是对商品标题而言的，Co_ratio_{click} 表示两词共现比例，$C(w_{i_t}w_{j_q})$ 表示词 w_i 在点击标题中且 w_j 在查询中的数目。MI_{click} 表示两词共现互信息，$p(w_{i_t}w_{j_q})$ 表示词 w_i 在点击标题中且 w_j 在查询中的频率，$p(w_i)$ 表示词 w_i 的频率。$Front\_ratio_{click}$ 表示两词互换比例，$Query\_ratio_{click}$ 表示两个词都出现在标题中时，查询中只出现词 w_1 与只出现词 w_2 的比例，$C_q(w_i)$ 表示查询中只出现 w_i 的数目。

3.3 Gradient Boosting Decision Tree 模型

Gradient Boosting Decision Tree（GBDT）模型是一种组合模型，它的基本思想是迭代的构建决策树，最后得到一个由 M 棵决策树组合而成的模型从而避免了单棵决策树过拟合的缺点[17]。

在同义词识别问题中，x 代表特征集合，y 代表相关性分数集合，(x_i, y_i) 代表每个词对，$P = \{\beta_m, \alpha_m\}_1^M$ 代表参数集合。给定特征集合 x 和参数集合 P，GBDT 模型由式(1)计算得到相关性分数。

$$F(x, P) = \sum_{m=1}^{M}\beta_m h(x, \alpha_m) \tag{12}$$

训练过程就是根据已知特征集合 x 和相关性分数集合 y，用式(2)求参数集合 P，即使得每个词对在模型 $F(x, P)$ 下的损失函数 $L(y, F(x, P))$ 最小。

$$\{\beta_m, \alpha_m\}_1^M = \arg\min = \sum_{i=1}^{N} L(y_i, \sum_{m=1}^{M}\beta_m h(x, \alpha_m)) \tag{13}$$

组合模型是多个简单模型的组合，但效果比单个复杂模型更好，这一优势使得越来越多的人青睐于组合模型。GBDT 被广泛应用于分类、回归、排序等机器学习问题之中，表现出特有的优势[18] [19] [20]。

4 实验与分析

4.1 数据

本实验使用某电子商务网站 280 万条商品标题，和点击到这些标题的 360 万个查询。共得到 150 万候选同义词对，对其中 3900 词对进行手工标注，对 GBDT 模型标注值为 0 或 1，对 SVM 的标注值为-1 或 1。将标注数据均分为四份：三份用作训练集，一份用作测试集。使用上文特征集合构造特征，分别训练和测试。GBDT 模型的相似度取阈值为 0.5，即大于等于此阈值为同义词，反之则不同义。

4.2 实验结果与分析

实验结果如表所示：

表1 实验结果

模型	类别	准确率	召回率	F值
GBDT	同义词	0.5446	0.2697	0.3606
	非同义词	0.8451	0.9464	0.8737
SVM	同义词	0.5078	0.1288	0.2048
	非同义词	0.8241	0.9702	0.8917

两个模型都没有采用使 F 值最高的参数，而是选择了使准确率较高的参数。因为应用到电子商务检索中的同义词表必须是绝对准确的，这样才能有效地扩展查询，提高检索精度。因此需要在此结果的基础上进行人工校验，出于对校验成本的考虑，我们更侧重于准确率。

分析实验结果可知，影响准确率的因素主要是两词大量共现或互相点击，而不是同义词，比如新娘-伴娘。影响召回的因素主要是数据稀疏导致特征得分低，比如阿童木-铁臂阿童木。

5 总结和展望

本文在充分研究电子商务中用词特点的基础上，提出基于卖方用户行为和买方用户行为的同义词识别方法。通过并列关系符号切分商品标题和基于 SimRank 思想聚集查询两种方法获取候选集合，获取字面特征及其标题、查询、点击等用户行为特征，采用 GBDT 模型对候选集合中的词进行判定。实验表明这种方法识别的准确率达到 54.46%。下一步将继续深入挖掘标题、查询、点击等用户行为相关的特征，以期达到更好的效果。

参 考 文 献

[1] H. Coleridge, J. Murray, H. Sweet et al. 2005. The Oxford English Dictionary. Oxford University Press. Oxford.

[2] N. Kanhabua, K.Norvag. Exploiting time-based synonyms in searching document archives. In *Proceedings of ECDL*. 2010.

[3] 宋明亮. 汉语词汇字面相似性原理与后控制词表动态维护研究. 情报学报. 1996, (04).

[4] 吴志强. 经济信息检索后控制词表的研究. 南京：南京农业大学硕士学位论文. 1999.

[5] 朱毅华. 智能搜索引擎中的同义词识别算法研究[D]. 南京：南京农业大学硕士学位论文. 2001.

[6] 穗志方, 俞士汶. 主题概念规范化研究中的自然语言处理策略.《第二届术语学、标准化与技术传播学术会议论文集》. 1998, pp. 367-374.

[7] 刘群, 李素建. 基于《知网》的词汇语义相似度计算. 中文计算语言学. 2002, 7(2): 59-76.

[8] V. D.Blondel, P. P.Senellart. Automatic extraction of synonyms in a dictionary. Technical Report 89, Universitecatholique de Louvain, Louvainla-neuve, 2001.Presented at the Text Mining Workshop 2002 in Arlington, Virginia.

[9] J. Jannink. Thesaurus entry extraction from an on-line dictionary. In Proceedings of Fusion '99, Sunnyvale CA, Jul 1999.

[10] H. Chen, K. J.Lynch. Automatic construction of networks of concepts characterizing document database. *IEEE Transactions on Systems, Man and Cybernetics*. 1992, 22(5): 885-902.

[11] G. Grefenstette. Automatic thesaurus generation from raw text using knowledge-poor techniques. In *Making Sense of Words. Ninth Annual Conference of the UW Centre for the New OED and text Research*. 1993, 9.

[12] P. D. Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In European Conference on Machine Learning. 2001, pp. 491-502.

[13] D. Higgins. Which statistic reflect semantics? Rethinking synonymy and word similarity. *International Conference on Linguistic Evidence*. 2004.

[14] 刘华梅, 侯汉清. 基于情报检索的汉语同义词识别初探. 理论与探索. 2005, 28(4), pp, 373-375.

[15] G. Jeh, J. Widom. Simrank: a measure of structural-context similarity. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 538-543, 2002.

[16] I. Antonellis, H. Garcia-Molina, Chi-Chao Chang, Simrank++: query rewriting through link analysis of the click graph, Proceedings of the VLDB Endowment, v.1 n.1, August 2008.

[17] J. H. Friedman, Greedy function approximation: a gradient boosting machine, *Ann. Statist.* 29(2001) (5), p. 1189.

[18] Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, Keke Chen: Cross-Market Model Adaptation with Pairwise Preference Data for Web Search Ranking. COLING (Posters) 2010: 18-26.

[19] Z. Zheng, K.Chen, G Sun, and H.Zha. 2007. A regression framework for learning ranking functions using relative relevance judgments. In SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp, 287-294.

[20] Bang Zhang, Getian Ye, Yang Wang, Jie Xu, Gunawan Herman: Finding shareable informative patterns and optimal coding matrix for multiclass boosting. ICCV 2009: 56-63.