# 商品品牌名称挖掘*

何正焱，王厚峰
北京大学 计算语言学教育部重点实验室，北京 100871
E-mail：{hezhengyan, wanghf}@pku.edu.cn

**摘　要**：百度百科包含了大量的实体和丰富的链接与分类关系，在中文领域含有大量人类知识。在商品品牌名称抽取的挖掘中，我们提出了发现新的品牌名称的基于图模型的半指导方法。利用百度百科中词条间的相关关系和开放分类，我们使用不同的准则计算词条间的相似度，结合词条和分类的关联性，分类与分类之间的关联性，使用标记传播算法，在1.3 兆词条上进行了品牌名称的挖掘。取得了较好的效果。

**关键词**：商品名挖掘；半监督学习；图算法

# Semi-supervised Method for Mining Product Name

He Zhengyan, Wang Houfeng
Key Laboratory of Computational Linguistics (Peking University) Ministry of Education, Beijing 100871
E-mail: {hezhengyan, wanghf}@pku.edu.cn

**Abstract:** Baidu baike contains large amount of human knowledge like named entity，numerous link relationship and category information. In order to recognize product or brand name from free text，we need list of known product name, this can be automatically extracted from encyclopedia. We propose a graph-based method to discover product name using a few seeds. We incorporate the "related entry" and "open category" structure of baidu baike to reinforce the similarity measure of each other, and perform a semi-supervised graph-based method on 1.3 million entries, and achieved satisfactory results.

**Keywords:** brand name mining; semi-supervised learning; graph method

## 1　引言

在中文命名体识别中，对识别人名、地名和机构名的研究较为深入。使用的方法主要有基于规则的命名体识别和基于序列标注的命名实体识别[1]。

商品和品牌名称的识别较人名、地名的识别较难。人名有一定的规律可循，且用字比较固定；地名相对变化不大。品牌名称的取名较随意，规律性不强，并且有很多来自外文译名，识别相对困难。

虽然命名实体在用字和上下文有一定规律，但命名实体识别通常是一个严重依赖人类知识的领域，在地名识别中经常用做特征的地名词典（gazetteer）[2]，机构名词典便是人类知识的体现。因此挖掘和收集机构名和商品名对命名体识别有很大作用，本文首先考虑收集和挖掘特定领域的品牌名称。

百度百科是一个较大的中文知识库，包含了大量的人物、地理、历史、机构、商业知识。我们使用少量的种子节点，利用百度百科固有的“开放分类”和“相关词条”构造词条间的相似度，使用半指导的方法扩充品牌名称列表。

## 2　相关工作

在一个链接丰富的图结构上定义相似度是一个被深入研究的领域 [3]。图上相似度度量的方法主要有基于图的如personalize pagerank, 其基本思想是将pagerank中某个节点的重启概率设置为1，静态分布后的排序就是其他节点对该节点的相似度。hitting time [4] 定义为从节点i随机游走在重

* 本文受国家自然科学基金资助（编号：60973053，91024009）和博士点基金（编号：20090001110047）资助。

新回到 i 之前到达 j 的期望步数，两个节点越相似，期望步数越小。Katz 得分[5]定义为节点 i 到节点 j 的长度为 k 的路径数的加权平均，加权系数随距离增加指数下降，当大多数权重集中在短路径上时，katz 得分类似于 common neighbors。公共邻节点（common neighbors）定义为两个节点共有的邻节点数，Adamic /Adar 得分[6]定义为公共邻节点的加权和，每个公共邻节点的权值是其度的对数值的倒数，其本质是对公共邻节点的改进。

在异质的图网络中，文献[7]在文章-作者的异构图网络中，利用作者间共同创作，文章间相互引用和作者和文章的写作关系，耦合两个 pagerank 的随机游走过程，同时对作者和文章排序。[8]提出了一种在任意异构图网络上计算相似度的框架，节点间的边含有类型和权值，权值可以通过在训练数据上的错误反向传播学习，相似度的计算结合了随机游走和重新排序（reranking）、随机游走历史（walk history）等信息，实际上相当于在不同类型的边上增加权重。

simfusion [9]使用分块矩阵分别表示同质节点和异质节点间的关系，混合后的矩阵称为 URM (unifled relationship matrix)，作者证实了不同类节点间的关系会增强同质节点间的相似度度量。实际相当于对不同类型的关系加上权值。其中 $\Sigma_j \lambda_{i,j} = 1, \lambda_{i,j} > 0$ 。

$$L_{URM} = \begin{pmatrix} \lambda_{11}L_{11} & \lambda_{12}L_{12} & \cdots & \lambda_{1n}L_{1n} \\ \lambda_{21}L_{21} & \lambda_{22}L_{22} & \cdots & \lambda_{2n}L_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{n1}L_{n1} & \lambda_{n2}L_{n2} & \cdots & \lambda_{nn}L_{nn} \end{pmatrix}$$

标记传播（1abel propagation）[10]是一种基于图的半监督的机器学习方法，相对于完全监督的学习算法，在较少训练数据的情况下具有较好的性能。标记传播中关键在于定义好转移矩阵 $T$，其中

$$T_{ij} = P(j \to i) = \frac{w_{ij}}{\Sigma_{k=1}^{l+u} w_{kj}}$$

$W_{ij}$ 是 $ij$ 的相似度。$T_{ij}$ 可以理解为 j 传递给 i 的相似度的难易程度。

# 3 品牌名称抽取

## 3.1 数据整理

我们从百度百科[1]收集了约 130 万个词条，从每个词条中提取出标题、别名（同义词跳转等）、开放分类、相关词条。开放分类不同于维基百科[2]的层次分类，倾向于扁平结构的标签（tag），命名较随意。因此虽然比较方便，却不够规范。比如一个词条可以是被标记为分类“中国地理”，另一个词条被标记为“地理”，虽然他们在概念上很接近，分类标记却不能匹配。这就造成了分类上的数据稀疏问题。

## 3.2 相似度表示

在本文中我们考虑两类信息“开放分类”（tag）和“相关词条”，而不考虑文档内容、文档结构、文档内链接、文档主题、作者协作编辑等信息。“相关词条”可以看作是类型相同的词条，具有相同“开放分类”的词条也视作相同类型的词条。130 万词条中有约 125 万包含至少一个开放分类，约 29 万个包含至少一个相关词条。

[1] http: // baike.baidu.com

[2] http: // en.wikipedia.org

相关词条间的等价关系相对准确，比如“北京大学”的相关词条包含科研院所和高校，基本属于同类实体；“舒服佳”的相关词条包含洗化用品品牌；但是这类信息相对较少。

“开放分类”信息较丰富，大多数的词条都包含开放分类信息，但是开放分类信息通常具有用词随意的特点，并且百度百科的分类体系不像 wikipedia 具有层次结构，而是类似于任意给定的标签。另一个现象是标记省略[11]，比如“张朝阳”的开放分类有“画家，教师，企业家”，却没有“人物”。因此需要处理分类（tag）之间的相似关系。

本文提出了类似 simfusion 中的相似度表示，结合上述两种信息，在给定少量种子的情况下，通过半指导的算法进行品牌名称的挖掘。

为了表述方便，定义一个词条 $i$ 的相关词条的集合为 $R(i)$，开放分类的集合为 $C(i)$；如果词条 $j \in R(i), j$ 是 $i$ 的邻节点。$N(i)$定义为邻节点的个数。

两个词条节点的相似度定义为它们公共邻节点的个数，

$$L_e(i,j) = |N(i) \cap N(j)|$$

词条和分类之间的关系定义为词条包含分类标签，

$$L_{ec}(i,j) = 1 \text{ if } j \in C(i)$$

分类与分类的相似度定义为它们在相同词条中共现的次数，实际是分类节点之间的公共词条节点个数。考虑到分类之间是具有层次结构和包含关系的，因此分类的相似度传播不是对称的。比如：P（人物|企业家）P（企业家|人物），由于“企业家”一定是“人物”，而“人物”未必是“企业家”，因此前者的概率要大于后者。

$$L_c(i,j) = L_c(j \to i) = P(i \mid j) = \frac{c(i,j)}{c(j)}$$

设同质节点和异质节点间相对重要性为 $\alpha$，总的相似度矩阵定义为：

$$L = \begin{pmatrix} \alpha L_e & (1-\alpha) L_{ec} \\ (1-\alpha) L_{ce} & \alpha L_c \end{pmatrix}$$

## 3.3 基于图的半指导学习算法

本文使用基于图的半指导学习算法，标记传播 (1abel propagation)[10]。其具体步骤如下：

1. 传递标记 $Y \leftarrow TY$
2. 对行归一化，即 $\forall i, \Sigma_c Y_{i,c} = 1$
3. 重置种子节点的概率分布 $Y$

$T$ 为相似度矩阵，对列做归一化，$T(i,j) = P(j \to i)$ 可以理解为 $j$ 传递标记给 $i$ 的难易程度。$l$、$u$ 分别为带标数据和不带标数据的个数，$C$ 为类别个数，$Y_{(l+u)\times C}$ 是所有数据在类别上的概率分布。

在这里我们设 $T=L$，如果不考虑节点的类别，实际上相当于将所有带标节点的标记不断传递给不带标数据，最后按照概率由高到低排序，获得与种子（认为是品牌的词条）的类别接近的词条或分类。

## 3.4 种子词条

我们手工设计了几十个不同领域的品牌名称（见表 1），包含日化、服装、汽车、电子、家电、餐饮、化妆品、食品等领域。由于品牌名称的定义广泛，可能包含几十种不同领域。每种领域内部链接通常丰富，分类较一致；类别之间链接相对较少，分类也相对分散。因此每个领域我们选择几个具有代表性的词条作为种子节点。

表1　品牌名称的种子节点

| |
|---|
| 舒肤佳 佳洁士 心相印 纳爱斯 雕牌 碧浪 杜蕾斯 / 阿玛尼 凡客诚品 美特斯邦威 百丽 李维斯 安踏 杰克琼斯 雅鹿 奥黛莉 ONLY / 兰博基尼 米其林 斯柯达 / 诺基亚 联想 佳能 宏碁 海尔 索尼 爱立信 希捷 华为 / 俏江南 东来顺 / 香奈儿 雅诗兰黛 卡地亚 美加净 宝洁 护舒宝 欧莱雅 / 雀巢 百事可乐 费列罗 麦斯威尔 星巴克 / 蒙牛 金丝猴 双汇 康师傅 / 五粮液 人头马 / 飞亚达 宝玑 万宝路 / 当当网 / 屈臣氏 / 百盛 来福士广场 / 华纳 滚石 |

# 4　实验与分析

## 4.1　实验设计和评价

我们从百度百科中收集了 130 万个词条进行实验。由于实验的数据量很大，矩阵运算我们使用 scipy[1] 的稀疏矩阵。我们过滤掉了不包含相关词条和开放分类的词条，过滤掉频率小于 5 的开放分类。利用 $L$ 作为相似度矩阵，经过标记传播算法迭代 1000 次，此时矩阵 $Y$ 每个元素的平均迭代误差小于 $10^{-4}$，可以认为基本收敛。

由于标记传播结果的概率分布 $Y$ 表明了某个词条和种子词条的相似性，我们将 120 万个词条按概率由高到低排列，得到词条列表。概率越大，排序越高，越可能是一个商品品牌名称。

由于收集的词条数目太多，我们还专门从 globrand[2] 搜集了 756 个品牌名称，其中 667 个在我们搜集的百科词条中或别名中存在。我们利用这 667 个词条在所有 120 万个词条中的 rank 值相加，相当于在所有正例中采样出 667 个样本点，以采样的 rank 均值作为所有正例的期望 rank 值。如果 rank 值越小，表明排名越靠前，模型效果越好。

定义 rank (e) 为词条 e 在所有 120 万个词条中的排序值，表 2 列出了不同 $\alpha$ 下 667 个样本词条的排序和。

表2　不同 $\alpha$ 下 667 个词条的排序和

| $\alpha$ | .3 | .5 | .5 | .7 | .9 | .999 |
|---|---|---|---|---|---|---|
| $\Sigma_e rank(e)$ | 27156787 | 26926982 | 26315549 | 26573952 | 35206104 | 38931623 |

从表 2 可以看出，当 $\alpha \to 1$ 时，逐渐忽略分类对词条的影响，相当于只考虑词条间的相似性，而不考虑类别对词条的影响，效果逐渐变差，这表明整合两种信息能够提高品牌名的 rank 值，产生更好的效果。

## 4.2　实验结果分析

我们人工检查了排序较高的非品牌词条。我们将其分为几类，见表 3。某些是由于包含的信息太少，而唯一包含的信息又与正例很相关，比如“板砖”，“掏耳勺”仅仅包含一个分类“日常”，而“日常”与很多洗化品牌相关；“苦事”的唯一一个相关词条“乐事”是品牌；另一些如“HR”、“名表”等虽然有多个分类和相关词条，但是仅有少数和品牌相关，即存在不一致性和多义性。如何建模这两种情况是我们将要考虑的方向。

在 667 个样本中，前 450 个排序都在 10000 以内。对 667 个品牌名称 rank 值较低的样例（表 4）进行分析，我们可以发现多数存在岐义和多义词现象，因此这类词条只在特定上下文下才是品牌名称（如：白云山，见表 4）。另外一些词条的“开放分类”或“相关词条”提供的信息太少，或使用了很少使用的分类名称；如何整合更多的文档结构和内容信息是另一个将要研究的方向。

[1] http: // www.scipy.org

[2] http: // www.globrand.com / brandlisttxt /

表3　排序较高的非品牌词条

| | |
|---|---|
| 板砖，掏耳勺 | 只有一个分类“日常”，是一个较常用的日化品牌的分类 |
| 苦事 | 含有相关词条“乐事”是一个品牌，因此该相关词条不准确 |
| HR，索尼 | 多义词，有一个义项是品牌 |
| 梅厚钧，名表，脱敏牙膏，海魂衫 | 某一个分类是常用品牌分类，或某一个相关词条是品牌，这类一般有多个分类 |

表4　排序较低的品牌名称。带有括号的是百度百科中未标成品牌名的词条，括号中是该词条实际所指企业和百度百科中的类别

| |
|---|
| 白云山(制药，地名)，哈佛(教育咨询，大学)，史密斯(aosmith 热水器，人名)，太平洋(保险，地理)，少林寺，甲骨文(公司，古文)，集美(家居，地名)，迪拜(酒店，地名)，亚细亚(陶瓷，地理)，达利(食品，画家)，东星航空，乐凯，嘉陵(摩托车，地理)，凯撒(意大利服饰，人物)，衡水老白干，SOHO 中国，华丰，新东方，罗氏(制药，姓氏)，永乐(电器，年号)，厦新，将进酒(白酒，文学)，AC 尼尔森，zippo，三九(制药，节气)，尚德(能源，学校)，UT 斯达康，加多宝，和其正，奇异王果，黄金酒，口子窖，长城电脑，友邦保险，雨润，奥妮，冠生园，三棵树，厦华，富力，迪比特，亚玛逊 |

## 5　结论

我们提出了一种基于图的半监督学习算法，从大量百科知识库中抽取品牌名称。结合百度百科的相关词条和开放分类两种链接关系，定义了结合两种关系的相似度表示方法，给定少量品牌领域的种子样例，挖掘出更多的品牌名称。实验中我们仅利用“开放分类”和“相关词条”两类信息，而没有利用诸如文档内容、文档结构、文档内链接、文档主题、作者协作编辑等信息，取得了较好的效果。使用我们的方法，可以在指定任意领域（如机构名作为种子）的少量实例的情况下，获取更多的该领域相关的概念。抽取出的词表可以用在命名实体识别领域。

下一步，我们将进一步利用和融合更多信息（如文档内容、文档内链接、文档模板结构等），并提出更合理和可行的评价方法。

## 参 考 文 献


[1] 周俊生，戴新宇，尹存燕，陈家骏. 基于层叠条件随机场模型的中文机构名自动识别. *电子学报*, 2006.

[2] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 2007.

[3] Purnamrita Sarkar. *Tractable Algorithms for Proximity Search on Large Graphs*. PhD thesis, Carnegie Mellon University, 2010.

[4] D. Aldous and J. Fill. *Reversible Markov Chains and Random Walks on Graphs*. Book in preparation.

[5] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 1953.

[6] Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 2003.

[7] Ding Zhou, Sergey A. Orshanskiy, Hongyuan Zha, and C. Lee Giles. Co-ranking authors and documents in a heterogeneous network. *Data Mining, IEEE International Conference on*, 0:739-744, 2007.

[8] Einat Minkov. *Adaptive Graph Walk Based Similarity Measures in Entity-Relation Graphs*. PhD thesis, Carnegie Mellon University, 2008.

[9] Wensi Xi, Edward A. Fox, Weiguo Fan, Benyu Zhang, Zheng Chen, Jun Yan, and Dong Zhuang. Simfusion: measuring similarity using unified relationship matrix. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'05, pages 130-137, New York, NY, USA, 2005. ACM.

[10] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, 2002.

[11] Xiance Si, Zhiyuan Liu, and Maosong Sun. Explore the structure of social tags by subsumption relations. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1011-1019, Beijing, China, August 2010. Coling 2010 Organizing Committee.