基于搜索引擎的关键词自动聚类法*)

邓健爽　郑启伦　彭　宏　邓维维

（华南理工大学计算机科学与工程学院人工智能实验室　广州 510641）

摘　要　互联网为用户提供了一个丰富的信息平台。然而，当前人们对互联网中海量信息的利用主要通过搜索引擎去查询相关的信息，互联网只是作为一个简单的信息库供用户检索。本文研究通过搜索引擎获得互联网信息并且在此基础上进行更高层次的知识挖掘——基于搜索引擎对关键词进行自动聚类。这是一个全新的研究，实验结果表明该方法具有理想的效果和新颖的构思。

关键词　知识搜索引擎，知识挖掘，聚类

Keywords Clustering Based on the Search Engine

DENG Jian-Shuang　ZHENG Qi-Lun　PENG Hong　DENG Wei-Wei

(Department of Computing Science, The South China University of Technology, Guangzhou 510641)

Abstract　Internet offers a great information platform for the users. However, people mainly use the search engines to query the large information in the Internet, so the Internet is only as the simple information database for the user's querying. This paper focuses on mining the higher-level knowledge from the information returned by the search engine—— keywords clustering based on the search engine, which is a novel research. The experiment shows that the way has perfect effect with novel conceive.

Keywords　Knowledge search engine, Knowledge mining, Clustering

1　引言

随着网络的发展，互联网上的信息呈指数快速增长。人们利用搜索引擎，例如 baidu，google[1] 从海量的信息中查找所需。然而，搜索引擎技术仅仅对互联网上的信息进行索引，根据用户提供的关键字查找相关的网页并返回给用户，没有能力去提供更进一步的信息。随着数据挖掘，人工智能等技术的发展，人们开始利用这些技术结合搜索引擎提供更加智能化的服务。例如：通过数据挖掘技术对网页进行自动分类[2~4]，对搜索引擎返回的结果进行自动聚类[5~6]，以及对搜索引擎提供人性化的导航检索服务[7~9]。然而，这些应用并没有改变搜索引擎的本质，只是为搜索引擎提供更加丰富和人性化的功能。当前的搜索引擎都只是提供一种信息检索，返回的结果网页都存在于互联网的某个具体的地方。本文提出了一个新的概念：知识搜索引擎。知识搜索引擎为用户提供的不是一种简单的信息查找服务，而是根据互联网中的信息为用户提供一种规律的查找和知识的查询服务。这种服务返回的信息不存在于互联网中某些具体的网页中，而是要综合整个互联网，对隐含的规律知识进行挖掘提取的结果。

知识搜索引擎包含一些具体的应用，例如在某个领域内对互联网相关知识的挖掘，发现一些规律，也可以根据互联网的丰富信息对一些概念进行自动理解和分类。本文主要讨论如何利用搜索引擎提供的互联网信息对一些名词或概念进行自动聚类。在机器学习领域，计算机要对不同的概念进行区分，必须学习大量的样例，同时需要一个学习过程，产生一个知识库，再对新的概念进行辨认。本文把互联网当成一个巨大的知识库，并且通过搜索引擎检索这个知识库对不同的名词/概念进行在线的分类。该方法不需要收集任何样例和漫长的学习过程，可以帮助用户辨别概念知识，也可为其他计算机程序提供一个智能化的信息预处理过程。

用户可以通过任意的查询关键词对搜索引擎进行查询，搜索引擎返回互联网上包含这些关键词的相关网页，并按相关度排序。经研究发现，一些主题相关的关键词存在大量的共同结果网页，并且这些网页都包含了它们共性的一些主题。当要对关键词"李宇春"，"何洁"，"刘德华"，"张学友"进行分类，计算机在没有任何先验知识的前提下是不可能完成该任务的。然而，互联网中存在着大量包含这些关键词的网页，我们猜想能否通过挖掘互联网中的信息，找到它们之间隐藏的联系进行自动聚类。本文通过搜索引擎检索这些关键词，得到它们相关的结果网页，并且分析这些网页之间的链接结构与文本信息得到关键词间的关联信息进而对关键词进行类别划分。

本文的第 2 节介绍影响关键词相关度的几个因素以及它们的计算公式，第 3 节介绍两种不同的聚类形式，第 4 节通过实验说明该方法的有效性和趣味性。

2　相关度的计算

搜索引擎包含的信息是互联网整体信息的一个缩影，通过这个缩影反映的关系来确定检索关键词间的联系。

图 1 中 A，B 分别为搜索引擎返回关键词 w_1，w_2 的相关

*)基金项目：广东省科技攻关项目(2005B10101033)，(A10202001)，广州市科技攻关项目(2004Z2-D0091)。**邓健爽**　博士研究生，主要研究方向为人工智能，网络智能搜索和数据挖掘。**郑启伦**　教授，博士生导师，主要研究方向为人工智能与海量数据处理相关的智能计算技术及其应用。**彭　宏**　教授，博士生导师，主要研究方向为数据挖掘。**邓维维**　博士研究生，主要研究方向为数据挖掘。

网页集，当相交的部分越大，则 w_1，w_2 越相关。

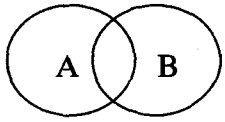

图 1　两关键词间结果网页的结构

定义 1(相交相关度 PR)

$$PR(w_1,w_2)=\left|\frac{A\cap B}{A\cup B}\right| \tag{1}$$

$Q(w_1,w_2,\cdots,w_n)$ 表示查询向量为 $w_1,w_2,\cdots,w_n$ 时搜索引擎返回的相关网页数，公式(1)表示为：

$$PR(w_1,w_2)=\frac{Q(w_1,w_2)}{Q(w_1)+Q(w_2)-Q(w_1,w_2)} \tag{2}$$

图 1 中 A,B 相交部分的网页在 A,B 各自的网页集中如果排名越前，则 A 和 B 的相关度越高。所以我们定义了另一个影响关键词之间相关度的因素。

定义 2(排名相关度 RR)　假设网页 p_i，且 $p_i\in A\cap B$，R_{Ai}，R_{Bi} 分别为网页 p_i 在网页结果集 A，B 中的排名，排名相关度的计算公式为：

$$\begin{aligned}&PR(w_1,w_2)\\&=\frac{1}{\frac{1}{2}\left(\frac{R_{A1}+R_{A2}+\cdots+R_{An}}{n}+\frac{R_{B1}+R_{B2}+\cdots+R_{Bn}}{n}\right)}\\&=\frac{2n}{\sum_{i=1}^{n}(R_{Ai}+R_{Bi})}\end{aligned} \tag{3}$$

其中 $n=|A\cap B|$。

搜索引擎返回的关键词查询结果由相关网页的标题以及包含查询关键词的网页文本信息组成，当中往往同时包含了与关键词主题相关的一些词语或短语，通过抽取这些频繁出现的相关词语可以得到查询关键词的一些主题信息。同时，越相类似或者越相关的查询关键词之间可能包含越多相同的主题词，所以通过查找相同主题词可以判断查询关键词之间的关联度。

定义 3(主题词相关度 SR)　设查询关键词 w_1 包含的主题词集合为 S_1，w_2 包含的主题词集合为 S_2，主题词相关度计算公式为：

$$SR(w_1,w_2)=\left|\frac{S_1\cap S_2}{S_1\cup S_2}\right| \tag{4}$$

由此得出两个关键词间的相关度计算公式为：

$$\begin{aligned}&Relating(w_1,w_2)\\&=k_1*PR(w_1,w_2)+k_2*PR(w_1,w_2)+k_3*SR(w_1,w_2)\end{aligned} \tag{5}$$

推导出 k 个关键词$\{w_1,w_2,\cdots,w_k\}$的相关计算公式为：

$$PR(w_1,w_2,\cdots,w_k)=\left|\frac{A_1\cap A_2\cap\cdots\cap A_k}{A_1\cup A_2\cup\cdots\cup A_k}\right| \tag{6}$$

$$RP(w_1,w_2,\cdots,w_k)=\frac{kn}{\sum_{i=1}^{n}(R_{A_1i}+R_{A_2i}+\cdots+R_{A_ki})} \tag{7}$$

$$SR(w_1,w_2,\cdots,w_k)=\left|\frac{S_1\cap S_2\cap\cdots\cap S_k}{S_1\cup S_2\cup\cdots\cup S_k}\right| \tag{8}$$

$$\begin{aligned}&Relating(w_1,\cdots,w_k)\\&=k_1*PR(w_1,\cdots,w_k)+k_2*PR(w_1,\cdots,w_k)+k_3*SR(w_1,\cdots,w_k)\end{aligned} \tag{9}$$

3　聚类形式选择

我们的方法根据用户输入的不同参数类型为用户提供两种不同的聚类形式。

3.1　输入相关度参数

候选相关集的选择：

判别多个(超过 2 个)关键词是否属于同一类，我们根据以下原则选定候选相关集：属于同一类的关键词集合内部的任何非空子集都是属于同一类的。假设有 A,B,C,D,E 5 个关键词，通过相关度计算公式得出它们间的相关矩阵如图 2。

	A	B	C	D	E
A	1	0.6	0.7	0.3	0.4
B	0.6	1	0.8	0.2	0.3
C	0.7	0.8	1	0.6	0.4
D	0.3	0.2	0.6	1	0.6
E	0.4	0.3	0.4	0.6	1

图 2　A,B,C,D,E 间相关矩阵

选定相关度大于 $\lambda=0.5$ 的为相关关键词，根据以上矩阵我们可以构造 2-相关类图：

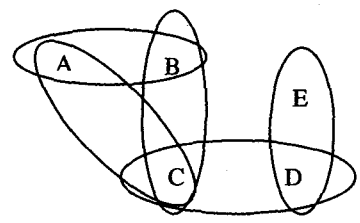

图 3　A,B,C,D,E 的 2-相关类图

图 3 中属于同一椭圆内的两个关键词为相关关键词，当要进一步查找 3-相关类(包含 3 个关键词的类别)时，只需要在上图中查找包含 3 个关键词的全连通子图，作为候选 3-相关类，再运用相关度计算公式计算该候选 3-相关类的相关度，大于 λ 则为 3-相关类。上例中得出候选 3-相关类为{A,B,C}，再计算 Relating(A,B,C)。

相关算法：

```
输入:关键词的 2-相关矩阵;相关度参数 λ
输出:各个极大相关类
步骤:
k=2;
do
{
  构造关键词的 k-相关类图;
  k++;
  while (发现候选 k-相关类)
    {
        计算该候选 k-相关类的相关度;
        if(候选 k-相关类的相关度)=相关度参数 λ)则
            输出该 k-相关类;
}
}while (有 k-相关类输出 && k(关键词数);
```

3.2　输入聚类数参数

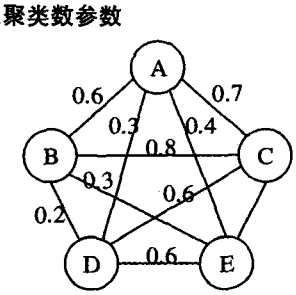

图 4　A,B,C,D,E 的相关连通图

当用户输入参数为聚类数的时候，根据图4，通过以下算法进行分裂聚类：

```
输入：整体相关连通图；聚类数 k
输出：k 个相关连通子图
步骤：
While (连通子图数目<聚类数 k)
{
    选择相关度最小的连接；
    断裂该连接；
}
```

4 实验

我们在Google的基础上进行了多组不同类型关键词的自动聚类实验，均取得了令人满意的效果。例如，对关键词组{李宇春，何洁，周笔畅，刘德华，张学友}，{香蕉，电视，葡萄，衣服，冰箱，鞋子}分别输入不同类型的参数进行聚类实验。

对每个关键词获取搜索引擎返回的前200条记录进行排名相关度以及主题词相关度的计算。相关度计算公式中取 $k_1=0,6, k_2=0,2, k_3=0,2$。

对关键词组{李宇春，何洁，周笔畅，刘德华，张学友}，我们输入相关度参数 $\lambda=0.08$，得出的结果为：

类别	关键词	部分主题词
1	李宇春，何洁，周笔畅	超级，超女，演出，活力，女生
2	刘德华，张学友	江湖，香港，黎明，郭富城，天王

对关键词组{香蕉，电视，葡萄，衣服，冰箱，鞋子}，输入聚类数 k=3 得出以下结果：

类别	关键词	主题词
1	香蕉，葡萄	果汁，橙汁，爵必思，金桔，木瓜，牛奶，哈密瓜，水果
2	电视，冰箱	行业，家电，价格，产品，洗衣机，空调
3	衣服，鞋子	店铺，服装，购物，女人，时尚，精品

实验结果表明，该方法能够有效地对关键词组进行自动聚类，不需要任何背景知识以及学习过程，可以方便地应用于其他计算机智能程序以及为用户提供了一些有趣的信息，使用户能够更进一步利用互联网中蕴含的丰富知识，为未来互联网的开发利用提供了一个突破口。

结论 当前人们利用互联网的信息主要是通过搜索引擎进行信息检索，这种方式只是对存在于互联网某一网站或某一网页中的信息进行快速的查找与简单的重现，并没有发现互联网中隐含的知识和规律。本文把当前的搜索引擎称之为信息搜索引擎并提出了一个新的概念：知识搜索引擎。知识搜索引擎通过对互联网信息的分析挖掘，产生更高层次的知识，这些知识不是静态的存在于互联网的某处，而是信息提炼升华的结果。本文通过一个具体的应用"基于搜索引擎的关键词自动聚类"来对知识搜索引擎的技术以及应用进行探索。基于搜索引擎的关键词自动聚类方法分析搜索引擎返回的关键词相关网页的链接结构以及文本信息，发现关键词间隐含的联系从而对关键词实现智能自动分组。实验结果表明该方法具有良好的效果，能够为其它计算机程序提供智能的预处理过程以及为用户提供更丰富和更有趣的知识。该方法使得用户可以更进一步利用互联网信息，是一个全新的研究。

研究知识搜索引擎其他的应用，例如通过搜索引擎，对互联网中具体某领域进行分析挖掘，发现某种产品或事件间的发展趋势以及规律。

参 考 文 献

1 Brin S, Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine

2 Liang Jiu-zhen. Chinese Web Page Classification Based on Self-Organizing Mapping Neural Networks. In: Proceedings of the Fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'03)

3 Shen Dou, Cong Yan, Sun Jian-ta, Lu W-chang. STUDIES ON CHINESE WEB PAGE CLASSIFICATION. In: Proceedings of the Second International Conference on Machine Learning and Cybernetics, Wan, November 2003

4 Yu H, Han Jiawei, Chang K C-C. PEBL: Web Page Classification without Negative Examples. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2004, 16(1)

5 张伟. 基于WWW的聚类搜索引擎的研究：[2003年全国优秀博士学位论文]

6 Zeng Hua-Jun, He Qi-Cai, Chen Zheng, Ma Wei-Ying. Learning To Cluster Search Results. In: The 27th Annual International ACM SIGIR Conference (SIGIR'2004), July 2004

7 Mobasher B, Cooley R, Jaideep S. Creating Adaptive Web Sites Through Usage-Based Clustering of URLs

8 Abawajy J H, Hu M. J. A new Internet meta-search engine and implementation. IEEE, 2005

9 Kawano H. Overview of Mondou web search engine using text mining and information visualizing technologies. IEEE, 2001

(上接第147页)

了刻画CA规则空间规律性的 δ 参数。本文结果表明，随着 δ 从0变化到1，CA规则空间发生从有序到混沌的相变。由于 $\delta=0.5$ 出现了混沌型CA，而复杂型和局部混沌型CA都出现在 $\delta=0.75$ 处，可以推测CA规则空间的相变包含多种途径，或者发生从有序到混沌的尖锐突变，或者相变过程中存在临界规则，这与朗顿等人提出的一阶相变和二阶相变一致；一阶相变和二阶相变的临界点分别发生在 $\delta=0.5$ 和 0.75 处，不存在关于临界值的争论，表明 δ 参数比 λ 参数能更好地刻画CA规则空间规律性和反映CA的本质特征。从应用的角度，δ 参数使得从巨大的CA规则空间中搜索适合特定应用问题的CA变得有规可寻。本文的研究表明，计算机实验和现代数学方法相结合研究CA是行之有效的方法。

当 r 值增大的时候，沃尔什谱密度的分布特征有待于进行更广泛的计算机实验和分析；对 $k>2$ 的CA，采用相同的基本思路，扩展二值的沃尔什变换到多值正交变换中，是值得进一步研究的问题。

参 考 文 献

1 Wolfram S. A new kind of science. USA: Wolfram Media Inc, 2002

2 Niloy G, Biplab K S, Andreas D, et al. A survey on cellular automata: [Technical report]. Centre for High Performance Computing, Dresden University of Technology, 2003. 1～28

3 Langton C G. Computation at the edge of chaos: Phase transitions and emergent computation. Physica D, 1990, 42: 12～37

4 Packard N H. Adaptation towards the edges of chaos. In: Kelso J A S, Mandell A J, Shlesinger M F, eds. Dynamic patterns in complex systems. Singapore: World Scientific, 1988. 293～301

5 Mitchell M, Hraber P T, Crutchfield J P. Revisiting the edges of chaos: Evolving cellular automata to perform computations. Complex Systems, 1993, 7(1): 89～130

6 Li Wentian, Packard N H, Langton C G. Transition phenomena in cellular automate rule space. Physica D, 1990, 45: 77～94

7 Li Wentian, Packard N H. The structure of the elementary cellular automata rule space. Complex Systems, 1990, 4(3): 281～298

8 Wootters W K, Langton C G. Is there a sharp phase transition for deterministic cellular automata? Physica D, 1990, 45: 95～104