中文自动分类在搜索引擎中的应用研究

李红宇　刘庆江

（哈尔滨师范大学阿城学院计算机与信息系　哈尔滨 150301）

摘　要　为了提高搜索引擎的查准率，帮助用户快速地定位其感兴趣的网页，可应用中文网页自动分类技术，实现快速准确的搜索引擎系统，使其具有较高的查准率。

关键词　中文自动分类，搜索引擎，Web挖掘

Application Research of Chinese Automatic Classification in Search Engine

LI Hong-yu　LIU Qing-jiang

(Department of Computer and information, A Cheng College, Haerbin Normal University, Haerbin 150301, China)

Abstract　To improve the precision of search engine and locate user-interesting Web page promptly, we apply Chinese Web page automatic categorization technology to search engines. Result shows this method has a high precision.

Keywords　Chinese automatic classification, Search engine, Web mining

1　引言

拥有海量网页信息的Web就像一本无所不包的百科全书。由于没有“主编”，人们可以随心所欲地向这本书提交任何信息，这样就导致了这本书在内容组织上的极端混乱。尽管它包含着极大的信息资源，但是真正有用的信息却相对匮乏。面对规模如此庞大的信息海洋，试图通过浏览Web网页来发现信息已经变得异常困难，往往花费大量的精力却所获甚少。因此，在Web用户和Web信息资源之间出现了巨大的鸿沟：一方面，Web资源中蕴含着极其丰富的有价值的信息和知识；另一方面，用户却无法有效地获取这些信息和知识。因此，为了能够有效地组织和分析海量的Web信息资源，帮助Web用户方便地获取其需要的信息和知识，人们希望能够按照其内容实现对网页的自动分类。

自动分类是用计算机系统代替人工对文献等对象进行分类，到Web信息检索、自然语言处理、机器学习等多个领域。一般包括自动聚类和自动归类。自动聚类指的是由计算机系统按照被考察对象的内部或者外部特征，按照一定的要求（如类别的数量限制，同类对象的亲近程度等等），将相近、相似或者相同特征的对象聚合在一起的过程。自动归类是指计算机系统按照一定的分类标准或者分类参考，将被考察对象划归到不同类目的过程[2]。

目前，很少搜索引擎提供对网页的分类浏览或检索，其原因之一是由人工进行网页的分类几乎是不可能的。如果能够实施网页的自动分类，就可以实现网页标引和检索的分类主题一体化，搜索引擎就能够兼有分类浏览、检索和关键词检索的优点，同时具备族性检索和特性检索的功能；能够深入到网页层次，帮助用户迅速地判断返回的结果是否符合自己的检索要求。例如在关键词检索中用熊猫作为检索词，返回的结果中作为动物的熊猫、作为一种杀毒软件的熊猫和作为一种电子产品的熊猫等内容是夹杂在一起的，用户要对结果进行分析判断，才能确定哪些是自己需要的。如果采用自动分类技术，就可将不同的内容分到不同的类目中去，从而节省用户判断时间，提高检索效率。本文探讨自动分类在搜索引擎中的应用。

2　自动分类的实现方法

2.1　自动归类的实现方法

目前，已有的主要文档自动分类算法可以分为三类：

（1）词匹配法。词匹配法可分简单词匹配法和基于同义词的词匹配法两种。简单词匹配法根据文档和类名中共同出现的词决定文档属于哪些类。该算法分类规则简单，但分类效果差。基于同义词的词匹配法是对简单词匹配法的改进，它先定义一张同义词表，然后根据文档和类名以及类的描述中共同出现的词决定文档属于哪些类。该算法扩大了词的匹配范围，性能上优于简单词匹配法。但算法的分类规则仍然机械，静态构成同义词表，对文档的上下文不敏感，无法正确处理文档中其具体含义依赖于上下文的词，分类的准确度很低。

（2）基于知识工程的方法。基于知识工程的文档分类方法，需要知识工程师手工地编制大量的推理规则，这些规则通常面向具体的领域，当处理不同领域的分类问题时，需要不同领域的专家制定不同的推理规则，而且分类质量严重依赖于推理规则的质量。因此，在实际的分类系统中较少使用基于知识工程的学习法。

（3）统计学习法。统计学习法和词匹配法在分类机制上有着本质的不同。它的基本思路是先收集一些与待分类文档同处一个领域的文档作为训练集，并由专家进行人工分类，保证分类的准确性，然后分析这些已经分好类的文档，从中挖掘关键词和类之间的联系，最后再利用这些学到的知识对文档分类，而不是机械地按词进行匹配。因此，这种方法通常忽略文档的语言学结构，而用关键词来表示文档，通过有指导的机器学习来训练分类器，最后利用训练过的分类器来对待分类的文档进行分类。这种基于统计的经验学习法由于具有较好

李红宇　硕士研究生，讲师，研究方向为Internet/Intranet应用软件；刘庆江　副教授，主要从事计算机体系结构及软件工程方面的研究工作。

的理论基础、简单的实现机制，以及较好的文档分类质量等优点，目前实用的分类系统基本上都是采用这种分类方法。文档自动分类算法的分类见图1。

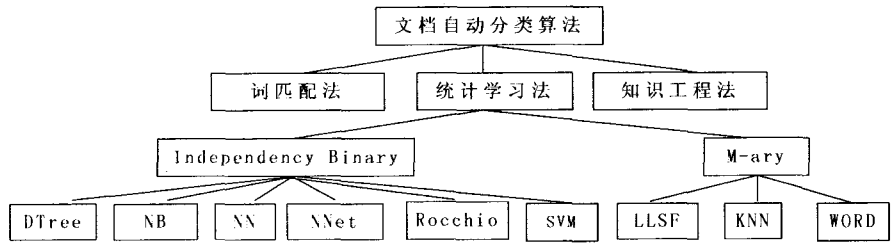

图1 文档自动分类算法的分类

本文介绍的文档分类是基于统计的自动分类系统，它忽略文本语言学结构，将文本作为特征项集合，利用加权特征项构成向量进行文本表示，利用词频信息对文本特征进行加权。实现起来简单，且分类准确度高，满足应用的要求。统计分类系统中广泛采用向量空间模型作为文本计算模型。向量空间模型可以将给定的文本转换成一个维数很高的向量。最突出的特点是可以方便地计算出两个向量的相似度，即向量所对应的文本的相似性。

在向量空间模型中，文本泛指各种机器可读的记录，用D(Document)表示，特征项(Term，用 t 表示)是指出现在文档 D 中且能够代表该文档内容的基本语言单位，主要是由词或者短语构成，文本可以用特征项集表示为 $D(T_1,T_2,\cdots,T_n)$，其中 T_k 是特征项，$1\leqslant k\leqslant N$。例如一篇文档中有 a,b,c,d 四个特征项，那么这篇文档就可以表示为 $D(a,b,c,d)$。对含有 n 个特征项的文本而言，通常会给每个特征项赋予一定的权重表示其重要程度。即 $D=D(T_1,W_1;T_2,W_2;\cdots,T_n,W_n)$，简记为 $D=D(W_1,W_2,\cdots,W_n)$，我们把它叫做文本 D 的向量表示。其中 W_k 是 T_k 的权重，$1\leqslant k\leqslant N$。在上面那个例子中，假设 a,b,c,d 的权重分别为30，20，20，10，那么该文本的向量表示为 $D(30,20,20,10)$。在向量空间模型中，两个文本 D_1 和 D_2 之间的内容相关度 $\mathrm{Sim}(D_1,D_2)$ 常用向量之间夹角的余弦值表示，公式为：

$$\mathrm{Sim}(D_1,D_2)=\cos\theta=\frac{\sum_{k=1}^{n}W_{1k}\times W_{2k}}{\sqrt{(\sum_{k=1}^{n}W_{1k}^2)(\sum_{k=1}^{n}W_{2k}^2)}}$$

其中，W_{1k}、W_{2k} 分别表示文本 D_1 和 D_2 第 K 个特征项的权值，$1\leqslant k\leqslant N$。

在自动归类中，可以利用类似的方法来计算待归类文档和某类目的相关度。例如文本 D_1 的特征项为 a,b,c,d，权值分别为30，20，20，10，类目 C_1 的特征项为 a,c,d,e，权值分别为40，30，20，10，则 D_1 的向量表示为 $D_1(30,20,20,10,0)$，C_1 的向量表示为 $C_1(40,0,30,20,10)$，则根据上式计算出来的文本 D_1 与类目 C_1 相关度是0.86。

网页自动归类一般包括以下步骤：

(1)网页特征的抽取和加权：网页特征的抽取是网页自动归类和自动聚类的前提。网页特征的抽取可以从以下几个方面提高网页自动分类系统的性能。首先是分类速度，通过网页特征的选择，可以大大减少特征集合中的特征数，从而提高网页自动归类系统的运行速度，使之能够满足现实需求。二是通过适当的特征选择，不但不会降低系统的准确性，反而会使系统的精度提高。这一点已经为实验所证明[3]。

为了使计算机能够更有效地处理网页特征，必须对网页特征进行特征加权，将网页特征表示成计算机能够处理的数学向量。网页数据是一种半结构化的数据，要比文本复杂得多。在网页表示中，对任一特征而言，有两个影响它权值的因素。一是该词的词频，另一个是该词在网页中出现的位置，在网页中不同位置出现的语词的价值是不同的。正如张琪玉教授指出："如果从针对文献整体的检准率的角度看，文献题名中的词最为有效。其次为文献中的小标题或者章节名、文献摘要。最后为文献中的词。"丁璇等人随机抽取了300篇经济类网页，对这些网页进行人工自由标引、人工打分、词频统计，并进行统计数据的分析、研究，得出了网页内容主题与网页题名、文章标题、第一段首句、第一段尾句、第二段首句、第二段尾句、第三段首句、第三段尾句、首段、尾段、HTML标记等12个标引源的主题表达能力的先后顺序。得出的结论是首段>文章标题>HTML标记>第一段首句>网页标题>第一段尾句>第二段首句>第二段尾句>尾段>第三段首句>其它>第三段尾句，并建议它们的加权值为5:5:5:4:4:4:2:2:2:2:2:2[4]。

(2)机器学习：机器学习的方法主要有支撑向量机(Support vector machine)、最近 K 邻居方法和贝叶斯算法等[5-9]。其中支撑向量机是建立在计算学习理论的结构风险最小化原则之上，其主要思想针对两类分类问题，在高维空间中寻找一个超平面作为两类的分割，以保证最小的分类错误率。支撑向量机的原理如图2所示，其中的实心点和空心点代表两个类别的训练样本，H 为将这两个类别分开的分类线，$H1$ 和 $H2$ 分别是经过这两个类别样本中距分类线最近的点且平行于分类线的直线，$H1$ 和 $H2$ 之间的距离叫做这两个类别的分类间隙。支撑向量机的目标是找到最优分类线，最优分类线不但能将两个类别的样本准确分开，而且使两类的分类间隙最大。

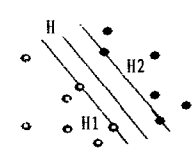

图2 支撑向量机的原理

最近 K 邻居方法的基本思路是在给定新网页后，考虑在训练网页集中与该网页距离最近(最相似)的 K 篇文本，根据这 K 篇网页所属的类别判断新网页所属的类别。它首先根据特征项集合来对训练网页向量重新描述，在新的网页达到首先确定新网页的向量表示，然后在训练网页中选出与新网页最相似的 K 个网页。也是根据网页的向量之间的距离，具体如下：

(下转第297页)

库的应用领域，在管理信息系统、专家系统、决策支持系统、群体工作环境系统中占有重要的位置。本文基于主动模糊数据库的原理，设计与实现 AFDB，并着重就模糊数据的定义与录入，模糊事件和模糊 ECA 进行了讨论。

参 考 文 献

[1] Bosc P, Pivert O. SQLf: a relational database language for fuzzy querying. IEEE Transactions on Fuzzy Systems, 1995. 3
[2] Peng T, Zuo W L, Liu Y L. Characterization of Evaluation Metrics in Topical Web Crawling Based on Generic Algorithm// 1th Intl Conf. ICNC, Changsha; Springer, LNCS. 3611. (SCI DBA23), 2005:690-697
[3] 左万利，刘居红. 关联图与主动规则集的终止性分析. 软件学报，2001(12):278-280
[4] 何新贵. 模糊关系型数据库的数据模型. 计算机学报，1989(2)
[5] 郝忠孝，熊中敏. 计算主动数据库中不可规约集合的有效算法. 计算机研究与发展，2006(1):285
[6] 关系型模糊数据库管理系统的设计与实现. 学位论文. 上海海事大学，2005,6
[7] 魏延. 主动模糊数据库中的事件与规则. 重庆师范学院学报，2002(12):175

（上接第 293 页）

$$\mathrm{Sim}(d_i,d_j)=\frac{\sum_{k=1}^{M}W_{ik}\times W_{jk}}{\sqrt{(\sum_{k=1}^{M}W_{ik}^2)(\sum_{k=1}^{M}W_{jk}^2)}}$$

K 值的确定是一个关键的问题。现在的一般做法是先选定一个初始值（几百到几千之间），在进行自动归类的过程中根据结果进行调整。接下来在新网页的 K 个邻居中，依次计算每一类的权重，计算公式为：

$$p(\vec{x},C_j)=\sum_{\vec{d}_i\in KNN}\mathrm{Sim}(\vec{x},\vec{d}_i)y(\vec{d}_i,C_j)$$

其中，$\vec{x}$ 为网页的特征向量，$\mathrm{Sim}(\vec{x},\vec{d}_i)$ 为相似度计算公式，而 $y(\vec{d}_i,C_j)$ 为类别属性函数，如果 $\vec{d}_i$ 属于类 C_j，那么函数值为 1，否则为 0。最后比较类的权重，将网页分到权重最大的那个类别中去。

2.2 网页自动聚类实现方法

网页的自动聚类一般包括四个步骤：

(1)网页表示：包括特征抽取和特征选择。特征选择是选择那些最具有区分性的特征，也就是最能把不同类别区分开来的特征，而不是大多数对象都具有的特征。

(2)相似度计算。主要根据网页表示的距离函数来定义。

(3)聚类：根据网页表示和相似度计算结果，按照规则将聚类网页分成不同的类。

(4)给出聚类的标识。在形成的每一类中抽取具有代表性的特征，作为该类的标识。

3 自动分类在搜索引擎中的应用策略

3.1 自动聚类和自动归类的应用

就目前的情况而言，自动聚类在搜索引擎中的实现要比自动归类容易一些，聚类的效果也比较显著。因此，可以考虑在搜索引擎中首先采用自动聚类。

如果要使用自动归类，首先就要考虑使用什么分类法。现在使用的分类法中既有传统的图书馆分类法，也有适应网络环境而生的网络分类法。二者各有千秋，传统的图书馆分类法系统性强，使用范围广，网络分类法比较灵活。如果条件许可的话，最好是两种类型的分类法都使用。对于熟悉图书馆分类法的用户就提供图书馆分类法的结果，对于一般用户则提供自编的网络分类法。在使用分类法的时侯，还要考虑分类的粗细问题，也就是分到几级类目。对于网页的分类，可能没有必要分得很细。下面主要论述自动聚类实现时涉及到的问题。

3.2 应用的时机

应用的时机是指自动聚类是在对网页数据进行索引的时候实施，还是在搜索引擎返回检索结果之后实施。前者可以利用网页的全文，后者一般只是使用网页的网址、标题和摘要等少量信息。一般而言，前者的结果要准确一些，但是综合考虑，后者的精确度虽然不如前者，但是成本比较低，实用性更强。它不需要对网页进行标引等预处理，工作量会大大降低，并且随着技术的发展，结果也会越来越令人满意。对于结果相关性的判断，既有客观因素，也有主观因素。机器只能够模拟人的思维而不能取代人的活动。自动聚类只是帮助用户进行相关性的判断而已，想靠它一劳永逸地解决相关性判断是不太现实的。

3.3 应用的对象

自动聚类可以应用到元搜索引擎或者单个搜索引擎中。单个搜索引擎的覆盖范围有限，且随着网络信息资源的迅速增长而不断下降，所以将自动分类应用于元搜索引擎返回的结果要比应用到单个搜索引擎的效果要明显一些。当然，元搜索引擎在对调用的搜索引擎进行选择时必须要遵循一定的原则，要选取质量比较高的，覆盖面比较广的，力争扩大检全率和检准率。对于单个搜索引擎返回结果，也没有必要全部包括在内，只需要前面的一部分就可以了（例如 50 条左右）。因为一般情况下，前面的结果与检索要求的相关度要高一些，这样做对于系统的精确性不会有太大程度的影响，但是可以将系统的成本大大降低，实用性更高。

3.4 用户界面

用户界面的设计是一个经常被忽略的问题，实际上用户界面的设计对于自动分类系统的使用效果有很大的影响。一个有关这方面的实验就证明了这一点。这个实验是 Hao Chen 和 Susan Dumais 完成的[20]。他们对七种检索界面的使用效果做了对比。这七种用户界面是：

(1)悬浮显示摘要的清单式界面，就是只有当鼠标移到返回的网页的标题时才显示出该网页内容的概要。

(2)内嵌摘要的清单是用户界面，将网页的摘要出现在返回网页的标题下面。

(3)显示类名的清单式界面，将返回网页的标题后面出现其所属的类目名称，同时给出网页的摘要。

(4)悬浮显示摘要的分类界面，首先给出类目的名称，然后显示出该类目下的网页标题，当鼠标移到该标题上的时候显示出该网页的摘要。

(5)内嵌显示摘要的分类界面，它与第四种界面基本上一样，除了是将网页的摘要显示在标题下面。

(6)无类名的分类界面，它将类目的名称和网页的摘要都去掉了。

(7)无网页标题的界面，只显示出类目供浏览。

结束语 综上所述，我们认为现阶段自动分类在搜索引擎中的应用主要应该考虑自动聚类在搜索引擎特别是元搜索引擎中的应用，将搜索引擎的结果进行自动聚类后返回给用户。采用类似于 Vivisimo 的用户界面，将类目的名称和网页的摘要明确地展现给用户，用户可以根据自动分类结果进行检索策略的修改。

参 考 文 献

[1] 冯是聪. 一种中文网页自动分类方法的实现及其应用. 计算机工程，2004(3):70-72
[2] 李晓黎. 基于向量机和无监督聚类相结合的中文网页分类器. 计算机学报，2001(1):62-68
[3] 秦兵，等. 可分性判据在中文网页分类中的应用. 微处理机，2002(1):26-28
[4] 范焱，等. 用 Naive Bayes 方法协调分类 Web 网页. 软件学报，2001(9):1386-1391
[5] Zamir O. A dynamic clustering interface to web search results// Eighth International World Wide Web Conference, 1999(5):11-14