第32卷 第24期 计 算 机 工 程 2006年12月
Vol.32 No.24 Computer Engineering December 2006
·人工智能及识别技术· 文章编号：1000—3428(2006)24—0206—03 文献标识码：A 中图分类号：TP319

CDSE：一个面向领域的智能搜索引擎

钟敏娟[1]，凌传繁[1]，白耀辉[2]，郭 攀[3]
(1. 江西财经大学信息管理学院，南昌 330013；2. 江西财经大学电子学院，南昌 330013；
3. 江西科技师范学院数学与计算机科学系，南昌 330013)

摘 要：介绍了一个面向领域的智能搜索引擎CDSE(Computer Document Search Engine)的设计和实现。CDSE结合文本分类和关键词组抽取检索用户需要的信息。利用了多个算法，综合运用了统计学方法、数据挖掘技术和Agent技术，较好地解决了现有搜索引擎普遍存在的搜索精度差、相关文档列序较后的问题。

关键词：智能搜索引擎；关键词组抽取；数据挖掘；Agent

CDSE：A Domain-based Intelligent Search Engine

ZHONG Minjuan[1], LING Chuanfan[1], BAI Yaohui[2], GUO Pan[3]
(1. College of Information Technology and Management, Jiangxi University of Finance and Economy, Nanchang 330013;
2. College of Electronics, Jiangxi University of Finance and Economy, Nanchang 330013;
3. Department of Mathematics and Computer Science, Jiangxi Technology & Science Normal College, Nanchang 330013)

【Abstract】CDSE, a model for domain-based intelligent search engine is proposed. The model can help people to retrieval what they need by combining text classification with key phrase extraction. Several algorithms that use key technology are proposed, such as statistics, data mining and agent. These algorithms solve the shortcomings effectively of low precision and relevant document ranking behind in now search engine.

【Key words】Intelligent search engine; Key phrase extraction; Data mining; Agent

当前Internet上包含了大量信息资源，Web已经拥有上100亿的静态网页[1]。对于目前通用的搜索引擎而言是不可能搜索到整个Web上的网页信息。在这些通用的搜索引擎中，各种引擎所搜索到的信息大部分是交叉的，而且内容繁杂、查询精度低、信息相对比较过时，从而导致了基于领域的智能搜索引擎发展。Justin Boyan提出运用机器学习机制优化搜索引擎的方法[2]，McCallum在此基础上运用机器学习机制建造了一个专门搜索计算机科学领域论文的专题搜索引擎Cora[3]。

参考以上文献资料，本文利用人工智能技术，特别是机器学习技术、多Agent技术，结合计算机领域相关论文的特点，设计实现了一个面向领域的智能搜索引擎CDSE。CDSE系统模型的主要特点如下：

(1)为了更好地提高检索精度，本文首先对用户提交的查询条件进行有效扩展。从而避免了词的不匹配现象导致一些相关文档不能被检索出来。

(2)充分利用文本分类和信息抽取技术形成反映类别文档内容的类关键词组集。它把机器学习技术、统计学方法、关联规则挖掘法结合起来，具有不需人工干预、准确率较高、自适应性较强的特点。

(3)利用信息检索技术将最终结果返回给用户。本文利用文档之间的引用关系以及对页面上的信息进行位置分析，从而使得最为相关文档排列在前。

1 CDSEngine系统模型

1.1 CDSE系统的体系结构

CDSE系统是一个多Agent系统，整个系统有较明显的层次关系，自顶向下依次分为2层：人机交互层和信息处理层。对CDSE系统模型中的2个层次进行进一步细化得到如图1所示的系统体系结构。

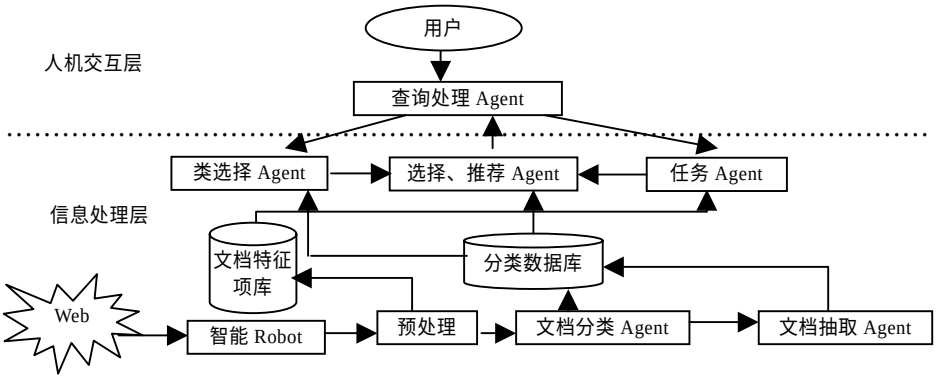

图1 CDSE系统模型

1.2 CDSE系统的功能

1.2.1 人机交互层

(1)用户：用户向系统发出查询请求和接受系统的服务。

基金项目：江西省教育厅科技基金资助项目(赣教计字[2005]326, 赣教计字[2005]327)

作者简介：钟敏娟(1976－)，女，讲师，主研方向：智能搜索引擎；凌传繁，教授；白耀辉，副教授；郭 攀，讲师

收稿日期：2006-02-02 **E-mail：**lucyzmj@sina.com

查询请求通常为关键词形式，系统将最为相关的结果返回给用户。

(2)查询处理 Agent：在实际的语言中存在着大量语义相关(包括同义、近义)和语用相关现象。如果完全以用户提交的关键词进行查询，则很可能会导致用户选择使用的词与文件集中出现的词不匹配，从而造成检索效果降低甚至失败。因此，必须对用户所提交的查询关键词进行扩展，从而最终产生系统的查询请求。

1.2.2 信息处理层

信息处理层 Agent 的具体功能如下：

(1)类选择 Agent：该 Agent 从查询处理 Agent 处接收查询关键词，计算这些关键词与分类数据库中每个类的关键词组集的相似度，找出相似度大于某一指定阈值的类。

(2)文档分类 Agent：文本分类 Agent 接收到预处理后的页面，对页面进行自动分类，建立分类目录树。分类器首先按照预先指定的标准对文档进行归类，然后利用机器学习机制对大量的文档进行快速有效地分类。

(3)文档抽取 Agent：对文档分类 Agent 返回的每一类文档，抽取每篇文档的头部信息(标题、摘要、关键词)和参考文献。利用关键词组抽取算法 extrkeyphrase 获得一组关键词组，然后根据文档所属类别构建类关键词集。

(4)选择、推荐 Agent：根据类选择 Agent 得出的类别信息，到分类数据库中相应类的文档特征项集中进行搜索。搜索时，分两种情况来进行，首先进行查询串与文档关键词组的匹配，如果匹配，则认为该篇文档是相关的，并赋以最大相似度值，否则利用 n_search 算法得出每篇论文与查询条件的相似度，并将相似度大于某一指定阈值的论文按照由大到小的次序返回给用户。

(5)任务 Agent：如果没有从类选择 Agent 中找出相对应的类，任务 Agent 就从查询处理 Agent 处接收已处理过的查询请求，到文档特征项库中进行相关信息的搜索。搜索策略仍然采用 n_search 算法进行。

2 CDSE 系统中关键技术的实现

2.1 构造类关键词组集

2.1.1 关键词组抽取算法 extrkeyphrase

该算法用来抽取单个文档的关键词集，主要是从一些频率较高的单词开始，根据每个单词的前后词找出相互关联的词组或短语，最终构成一组关键词。以前的研究人员总是处理整篇文档，但本文只是抽取文档的头部信息来提取关键词。具体算法描述为：

输入：单个文档的头部信息

输出：关键词集 keyphraseset

(1)对头部信息中每个单词(助词，虚词，副词剔除)进行词频统计并排序，找出前 n 个频率最高的单词，存放到 set 集合里；

(2)对 keyphraseset，number，maxwordnum 进行初始化，使得 keyphraseset=NIL，number=1，maxwordnum=m，其中 m 表示一个关键词组最多包含的单词个数；

(3)取出 set 集合中的第一个元素，words=first(set)；

(4)从 word 出发，分别和上文中前 number 个词，下文中后 number 词组成词组，存放在集合 temphrase 里；

(5)取 set 中的下个元素 word，当 word 不为空时转步骤(5)；

(6)对 n 进行调整，使得 n=n-number；

(7)计算 temphrase 中频率最高的 n 个词组，并将结果存放在 set 以及 keyphraseset 集合里；

(8)调整 number 的值，使 number=number+1，当 number 的值小于 maxwordnum 时，转(4)。

2.1.2 类关键词组权值计算算法

类关键词组权值意味着关键词组与每类文档的相似程度。权值越大，表明相似程度越高，对每类的贡献程度越大。通常，关键词组的相关度权值要考虑来自两方面的贡献[4]：局部权值 $L(i, j)$ 和全局权值 $C(i)$。这样有：

$$L(i, j) = \sum_{i=1}^{n} tf\_keyphrase_{ji} \tag{1}$$

$$C(i) = idf_k = \log_2(\frac{N}{n_k}) + 1 \tag{2}$$

其中，$tf\_keyphrase_{ji}$ 表示第 j 个关键词组在第 i 类文档中出现的频度，N 表示整个数据集划分的类别数目，n_k 表示整个类别数目中有多少个类出现了该关键词组。

2.2 选择相关论文算法 n_search

文献[5,6,7]提出同一关键词在 Web 文档中不同位置时所表达文档内容的能力是有差别的，如出现在标题中的特征项要比出现在摘要中的特征项更能确切代表文档的内容。文献[8]提出论文之间的相互引用很大程度上反映了彼此内容的相关性。n_search 算法综合考虑了这些因素，它首先将查询关键词与抽取出来的参考文献中论文标题进行内容匹配，如果匹配，则对链宿论文赋以标记，并计算该论文的权值 weight；否则对该论文部分内容(正文标题、摘要、正文引文)进行扫描，也计算出相应的权值 weight。

权值 weigh 计算算法，针对两种情况来进行。

(1)当查询条件与抽取出来的参考文献中论文标题相匹配时，根据N层向量空间模型算法[6]一篇文档可以从逻辑上划分为N个相对独立的文本段。参考文献作为其中一个独立的文本段，可以利用此算法中权值的计算公式得出每篇文档的权值，公式如下：

$$W_{ik} = tf_{ik} / L_i \tag{3}$$

其中，W_{ik} 表示特征项 t_k 代表文本段 S_i 的能力大小，L_i 表示文本段 S_i 长度，tf_{ik} 表示特征项 t_k 在文本段 S_i 中出现的次数。则一条匹配的参考文献的权值为

$$W_j = \sum_{k=1}^{m} W_{ik} \tag{4}$$

其中 m 为查询条件中不同特征项的个数。整篇论文的权值为

$$weight = \sum_{j=1}^{l} W_j \tag{5}$$

其中 l 为匹配的参考文献条数。

(2)当查询条件与参考文献中论文标题不匹配时，则对正文标题、摘要和引文部分进行扫描。结合式(4)获得被匹配的文本段 S_i 与查询文本 QS 的相似度：

$$Sim(QS, S_i) = \cos\theta_i = \frac{\sum_{j=1}^{k} W_{ij} * q_j}{\sqrt{\left(\sum_{j=1}^{k} W_{ij}^2\right) * \left(\sum_{j=1}^{k} q_j^2\right)}}$$

$$i \in [1, 2...N] \quad N=3 \tag{6}$$

则整篇文档 d_l 的权值为：

$$weight = Sim(QS, d_i) = \sum_{j=1}^{N} Sim(QS, S_{ij}) / N \tag{7}$$

3 算法性能分析

extrkeyphrase 算法精度不仅关系到对每类文档内容描述的准确性，而且还影响着相关类别的判定。不同的参数取值具有不同的精度，本文针对文档大小以及实际情况，通过实验比较了 6 组不同参数取值情况下的词组正确率，结果如表 1 所示。

表 1 每类文档抽取关键词组的准确率

Class name	Documents Size(KB)	n=12 m=2	n=12 m=3	n=10 m=2	n=10 m=3	n=8 m=2	n=8 m=3
Agent&AI	78	0.796	0.782	0.811	0.817	0.805	0.847
Application& Application& Sotfware Engineer	86.3	0.781	0.763	0.792	0.804	0.808	0.837
Architecture	92.7	0.813	0.776	0.820	0.801	0.792	0.856
Compress	76.3	0.726	0.701	0.748	0.736	0.775	0.821
Database	83.9	0.742	0.706	0.753	0.749	0.781	0.828
IR & WWW	77.5	0.735	0.722	0.746	0.734	0.762	0.812
Maching learning	80.7	0.743	0.702	0.715	0.700	0.751	0.806
Network	93.8	0.789	0.771	0.759	0.765	0.774	0.816
Operating systems	87	0.717	0.705	0.755	0.745	0.793	0.838
Theory& Security	72.3	0.720	0.714	0.741	0.752	0.758	0.822
Average Precision		0.756	0.734	0.764	0.760	0.780	0.828

从表 1 可以看出当 m＝3，n＝8 时，extrkeyphrase 算法的平均精度最高。在后续的实验中取 m＝3，n＝8。

使用 n_search 算法对文档进行测试，测试结果见表 2。

表 2 使用 n_search 算法的实验结果

实验项目 / 实验结果	类别判断成功	类别判断不成功
测试文档数目	80	800
查询时间	1.425s	27.047s
结果准确的文档数目	42	52
查准率	0.5	0.46

从实验结果可以看出，类别判断成功可以减少查询时间，提高查准率。这主要是因为对文档进行了分类的效果。文本分类缩小了检索范围，使得检索时不需要对整个数据集进行匹配，而只要在最相关类中进行查找，从而有效地减少了检索时间。同时，通过文本分类还可以将主题思想接近的文档划分在一起，这在一定程度上首先对数据集进行了一道预筛选，使得检索时能够在与查询请求相对接近的范围内进行。

4 结束语

本文设计和实现了一个面向领域的智能搜索引擎，介绍了该系统的体系结构和各组成部分的设计思想，着重讨论了 3 种智能 Agent 的功能以及相应算法。本文设计实现的系统较好地解决了现有搜索引擎普遍存在的搜索精度差、相关文档列序较后的问题，并突出了面向某一领域进行检索的特色。

参考文献

1 Menezer F, Pant G, Srinivasan P. Topic-driven Crawlers: Machine Learning Issues[Z]. 2002. http://dollar.biz.uiowa.edu/~fil/Papers/TOIT.pdf.

2 Boyan J, Freitag D, Joachims T. A Machine Learning Architecture for Optimizing Web Search Engines[C]. Proc. of AAAI Workshop on Internet- based Information Systems, 1996.

3 McCallum A, Nigam K, Rennie J, et al. Building Domain-specific Search Engines with Machine Learning Techniques[C]. Working Notes of the AAAI Spring Symposium on Intelligent Agents in Cyberspace, 1999.

4 Husbands P, Simon H, Ding C. On the Use of Singular Value Decomposition for Text Retrieval[Z]. 2000. http://www.citeseer.nj.nec.com/540137.html.

5 Cutler M, Shib Y, Meng Weiyi. Using the Structure of HTML Documents to Improve Retrieval[C]. Proc. of USENIX Symposium on Internet Technologies and Systems, Monterey, California, 1997: 241-251.

6 陈治平, 林亚平, 童调生. 基于 N 层向量空间模型的信息检索算法[J]. 计算机研究与发展, 2002, 39(10): 1233-1237.

7 刘　芳, 卢正鼎. 有效地检索 HTML 文档[J]. 小型微型计算机系统, 2000, 21(9): 986-988.

8 Menczer F, Pant G, Srinivasan P. Evaluating Topic Driven Web Crawlers[C]. Proc. of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001: 241-249.

(上接第 182 页)

束条件集。利用矩阵推移变换对约束集进行处理，并在种群初始化时采用前置矩阵降阶的方法可得问题的合法解，有效解决了多约束条件问题。

(2)适应度函数中的参数 x1、x2 可根据实际需求进行相应调整，通过改变遗传算法的进化目标最终得到不同目标下的最优方案，目前研究考虑了成本和工期，后续还可引入过程风险等因素。

(3)算法运行初期收敛性较好，达到一定周期后收敛性有所下降。一方面与算法参数设置有关，另一方面与算子结构有关。可以考虑引进基于浓度选择机制和高斯变异等方法改进遗传算法的算子结构。

参考文献

1 IEEE Std 610.12-1990. IEEE Standard Glossary of Software Engineering Terminology[S]. 1991-02.

2 陈迎欣. 软件过程中活动规划与资源分配方法的研究[D]. 哈尔滨: 哈尔滨工程大学, 2003.

3 郑　超, 高连生. 蚁群算法在资源受限项目调度问题中的应用[J]. 计算机工程与应用, 2005, 41(27): 205-208.

4 韩万江, 姜立新. 软件项目管理案例教程[M]. 北京: 机械工业出版社, 2005: 86-88.

5 汪应洛. 系统工程理论、方法与应用(第 2 版)[M]. 北京: 高等教育出版社, 1992: 38-39.

6 邱模波. 软件过程管理及其环境研究[D]. 南京: 南京航空航天大学, 2003.

7 曹　哲, 高　诚. 软件工程[M]. 北京: 中国水利水电出版社, 2004: 21-22.

8 许国志. 系统科学[M]. 上海: 上海科技教育出版社, 2000: 22-23.

9 尹文君, 刘　民, 吴　澄. 带工艺约束并行机调度问题的一种新的遗传算法[J]. 电子学报, 2001, 29(11): 1483-1484.