基于 Lucene 和 LSA 的法律咨询系统①

尹芝芳, 王 鑫, 蔡文正, 李 鹤, 阮玲玲

(桂林电子科技大学 计算机科学与工程学院, 桂林 541004)

摘 要: 本文设计的法律咨询系统, 结合法律行业的现状, 以中文问答系统为原型, 结合了开源数据检索项目 Lucene.net, 扩展了数据的存储类型. 本文借助中科院研发的中文分词系统, 集成到 Lucene.Net 平台上, 弥补了其分词不足. 并使用互信息技术, 使同义的法律相关词语优先进行检索. 在中文问答系统的答案提取时, 经常出现答案的“漏取”和“错取”的情况, 本文提出了一种基于潜在语义分析(LSA)的问题和答案句子相似度计算方法, 利用空间向量模型作为表示方法, 借助潜在语义分析理论, 通过奇异值分解的降维方法构建了一个低维的语义空间, 并在语义空间上实现了问题与答案句子相似度计算. 经试验证明, 本系统具有较精准的查询正确率以及较少的运行计算时间.

关键词: Lucene.Net; LSA; 问答系统; 互信息

Law Consultation System Based on Lucene and LSA

YIN Zhi-Fang, WANG Xin, CAI Wen-Zheng, LI He, RUAN Ling-Ling

(College of Computer Science and Engineering, Guilin University of Electronic and Technology, Guilin 541004, China)

Abstract: The designation of this law consultation system, not only considers the situation of the legal profession and based on Chinese Question-Answering System as prototype, but also use searching technology Lucene.net which is a open source project that can preform on many kind of types file. This article also uses ICTCLAS and applies it to the Lucene that makes up for Lucene's lack of word segmentation and mutual information technology to make the law word to be priority search. This paper proposes a method to calculate similarity between question and sentence based on Latent Semantic Analysis (LSA). This method represents the question and sentence with space vector model, under the help of latent semantic analysis theory, and constructs a semantic space, which gets rids of the correlativity between word. And then similarity calculation between question and sentence is implemented in this semantic space. Experiments show that this system has the precision of the operation of the inquiry accuracy and less computation time.

Key words: Lucene.Net; LSA; Question-Answering system; mutual information

随着计算机的普及以及自然语言处理技术的深入研究, 智能问答技术已经发展到可以替代人工一部分工作的地步了. 著名的项目有上个世纪 60 年代研制的 LUNAR 系统, 专门回答有关阿波罗登月返回的月球岩石样本的地质分析问题. 随着机器学习、自然语言处理技术的发展, 问答系统也随之有了广阔的发展空间[1]. 如今, 问答系统已经应用到了各种行业的信息管理系统中, 例如教育系统、酒店服务系统等. 问答系统可以方便给客户提供其想了解的信息, 且该信息为大多数用户所关心. 然而, 问题系统在法律咨询行业的应用目前还处于初级阶段, 通用的问答系统无法适用于现今的法律行业[2]. 因此, 本文根据现状, 结合自然语言处理, 机器学习等知识, 旨在构建一个法律咨询系统, 给用户提供更好的咨询和帮助.

本文所研究的法律咨询系统的知识库分为问答库和案例库, 主要以数据库和文件形式进行存储. 问答

① 基金项目:国家自然科学基金(61262074)

收稿时间:2013-08-30;收到修改稿时间:2013-10-04

库主要存储一些用户提问的简明扼要的回答信息, 而案例库中则存储大量详细的法律案例. 当用户无法在问答库中搜索到相应的答案时, 系统会自动在案例库进行搜索, 为用户列出相近的案例, 保证信息始终有反馈. 系统的主要框架基于 ASP.NET 的 B/S 结构, 使用二次开发的 Lucene.Net 作为搜索引擎[3][4], IIS 作为系统服务器. 通过对问句进行分词、去停用词、扩展、计算相似度等步骤, 查找相关度最高的信息.

本文实现的咨询系统, 具有检索速度快、精确度高、系统设计合理等特点, 具有较好的用户体验.

1 整体框架

本系统的运行环境为 dotNetFramework 4.0 , 采用 ASP.NET MVC 框架, 服务器为 IIS6.0.

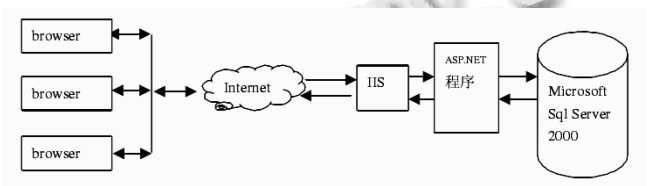

图 1 系统结构

系统需要调用大量数据如问题、答案、关键词信息等重要数据, 要求系统有较高的可靠性和安全性, 所以本系统选择了高质量、高性能的 SQL Server 2008 作为数据库服务器. 并使用 IIS 6.0 作为自动答疑系统的 Web Server..

本系统主要使用 ASP.NET 技术结合.NET 的组件技术来开发系统的智能咨询模块, 并应用 Lucene.Net 结合 ASP.NET 进行中文全文检索的二次开发.

系统采用 B/S 构架, 即 Browser/Server(浏览器/服务器)结构. 用户端使用浏览器, 即可登录到互联网上咨询.

2 知识库存储

本系统所涉及的问题答案和案例分别存储在问题库和案例库中. 答案库中分别存储着对应关系为一对一的问句和答案句子. 案例库则按照已有的分类进行信息的存储, 分类包括民法类, 商法类, 经济法类, 行政法类, 刑事法类等. 案例库中记录着每一篇案例的详细过程, 例如事件描述、律师分析等.

本文的检索功能使用的是 Lucene.net[5], 它主要包含了索引和搜索两个部分, 索引是主要建立不同文档的索引文件, 即对不同类型的文档进行处理, 包括 PDF、MSWord、HTML、Text File 等格式[6]. 因此, 答案库与案例库不但可以存储在数据库中, 还可以通过以上格式的文件进行存储. 多种格式的存储, 降低了数据搜集的难度, 可以快速建立庞大的知识库.

3 业务流程

该系统主要有两个登录身份: 普通会员和律师. 普通会员可以登录到该系统进行提问, 系统会给出贴近的答案. 如果系统不能满意的回答用户的提问, 那么系统将存储该问句, 标示为未解答, 并根据用户提出的问句, 检索案例库中的相关案例, 返回给用户以供参考. 律师登录到系统后, 会看到知识库中标示为未解答的问句, 并对其进行解答.

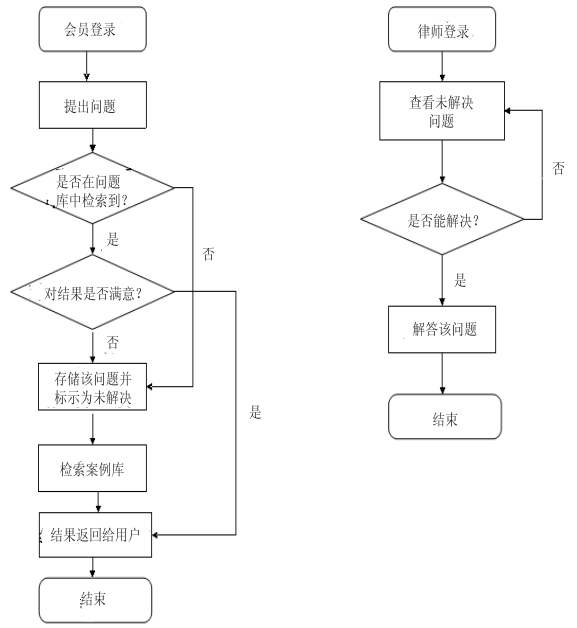

图 2 系统流程

4 问句处理

4.1 分词

由于汉语的特殊性, 不能像英文那样通过空格进行分词, 所以汉语的分词成为了近几年非常热门的研究. 主要方法有基于字符串匹配的方法, 例如正向最大匹配、逆向最大匹配、逐词匹配以及基于统计学的方法. 无论哪种方法都有其本身的优势和无法避免的劣势.

中国科学院计算技术研究所研制出的汉语词法分析系统 ICTCLAS, 是迄今为止最好的中文分词系统. 主要功能包括中文分词、词性标注、命名实体识别、新

词识别, 同时支持用户词典、支持繁体中文、支持 GBK、UTF-8 、 UTF-7 、 UNICODE 等多种编码格式. ICTCLAS3.0 分词速度单机 996KB/s, 分词精度 98.45%, API 不超过 200KB, 各种词典数据压缩后不到 3M.

4.2 去停用词

汉语中有很大一部分介词、代词、连词对文本内容影响不大, 对于这样的词, 系统将其放在一个称为停用词表(stoplist)的文本中. 文献[7]提出了一种新的停用词选取方法, 用该方法分别计算词条在语料库中各个句子内发生的概率和包含该词条的句子在语料库中的概率,在此基础上计算它们的联合熵,依据联合熵选取停用词.

4.3 基于互信息的扩展

互信息衡量的是某个词和类别之间的统计独立关系, 某个词 t 和某个类别 Ci 传统的互信息定义为:

$$\mathrm{MI(t,Ci)}=\log\frac{\mathrm{P}(t\wedge Ci)}{P(t)P(Ci)}=\log\frac{P(\mathrm{t|Ci})}{\mathrm{P(t)}}$$

$$=\log\frac{\mathrm{A}\times\mathrm{N}}{(A+C)(A+B)}$$

其中, A 为 t 和 Ci 同时出现的次数; B 为 t 出现而 Ci 没有出现的次数; C 为 Ci 出现而 t 没有出现的次数; N 为所有文档数. 如果 t 和 Ci 不相关, 则 MI(t,Ci)值为 0. 如果有 m 个类, 于是对于每个 t 会有 m 个值, 取它们的平均, 就可得到特征选取所需的一个线性序, 而 MI 平均值大的特征被选取的可能性大.

本文在进行特征词语抽取时, 使用了基于互信息的评估函数的方法, 对某些关键的法律术语之后出现的词条要加大评估值. 例如: 审判、诉称、依照等. 这样可以把法律术语优先提取出来, 增大不同类别案情文本的区分度, 从而提高查询的精确度.

4.4 基于潜在语义分析的相似度计算

潜在语义分析算法主要通过奇异值分解的方法对文档中语义结构进行计算并保留文本与词汇间最主要的关系. LSA 算法的步骤如下:

(1) 构造词一文档矩阵:

$$\mathrm{N}=\left[x_{i,j}\right]_{m\times n}$$

矩阵的行表示词, 列表示文档. $x_{i,j}$ 表示第 i 个词在第 j 个文档中的权重值.

TF 表示一个词在某一文档中出现的频率. IDF 由总文档数目除以包含该词语文档的数目, 再将得到的商取对数得到, 代表的含义是如果一个词在很多文档中均出现了, 则用它的识别度就会很低. 具体公式为:

$$\mathrm{idf}_i=\log\frac{\left|\mathrm{D}\right|}{1+\left|\left\{\mathrm{d}:\mathrm{t}_\mathrm{i}\in\mathrm{d}\right\}\right|}$$

|D|表示语料库中文章总数, $\left\{\mathrm{d}:\mathrm{t}_\mathrm{i}\in\mathrm{d}\right\}$ 是包含词语 $\mathrm{t_i}$ 的文档数目[8]. 则一个词的 TF-IDF 值计算公式为:

$$\mathrm{tfidf}_{\mathrm{i,j}}=\mathrm{tf}_{\mathrm{i,f}}\times\mathrm{idf}_\mathrm{i}$$

(2) 通过奇异值分解对矩阵 N 进行分解. 设 m≤n, rank(N)=r, 则 N 的奇异值分解记为: svd(N), 定义为

$$\mathrm{N}=\mathrm{U}\sum\mathrm{V}^\mathrm{T}$$

其中, U 和 V 均为酉矩阵(Unitary Matrix), 满足 $\mathrm{U^TU}=\mathrm{UU^T}=\mathrm{I_m}$, $\mathrm{V^TV}=\mathrm{VV^T}=\mathrm{I_n}$, $\mathrm{I_m}$ 和 $\mathrm{I_n}$ 分别为 m 阶和 n 阶单位矩阵. Σ 是含 N 所有奇异值的对角矩阵, 即 $\sum=\mathrm{diag}(\delta_1,\delta_2,\cdots\delta_r)$, $\delta_1,\delta_2,\cdots\delta_r$ 是 N 的 r 个奇异值且有 $\delta_1\geq\delta_2\geq\ldots\geq\delta_r$. U 和 V 的列向量分别是矩阵 N 的左右奇异向量.

(3) 对 SVD 分解后的矩阵进行降维, 求出矩阵的低阶近似.

矩阵的 SVD 分解具有以下性质:

设 $\mathrm{r}=\mathrm{rank}(\mathrm{N})\leq\min(\mathrm{m},\mathrm{n})$, $\mathrm{U}=(\mathrm{u}_1,\mathrm{u}_2,\ldots\mathrm{u}_m)$, $\mathrm{V}=(\mathrm{v}_1,\mathrm{v}_2,\ldots\mathrm{v}_\mathrm{n})$, 对于任一 0<k<r, 记为:

$$\mathrm{N}_k=\mathrm{U}_k\sum\nolimits_\mathrm{k}\mathrm{V}_\mathrm{k}^\mathrm{T}$$

其中 $\mathrm{U_k}=(\mathrm{u}_1,\mathrm{u}_2,\ldots\mathrm{u}_\mathrm{k})$, $\mathrm{V_k}=(\mathrm{v}_1,\mathrm{v}_2,\ldots\mathrm{v}_\mathrm{k})$, $\sum\nolimits_\mathrm{k}=\mathrm{diag}(\delta_1,\delta_2,\ldots\delta_\mathrm{k})$,

设 X 为 N 与之间的差, 即 X=N-$\mathrm{N_k}$, 则 X 的 F-范数为:

$$\left\|\mathrm{X}\right\|_\mathrm{F}=\left\|\mathrm{N}-\mathrm{N_k}\right\|_\mathrm{F}=\sqrt{\sum_{\mathrm{i=1}}^{\mathrm{m}}\sum_{\mathrm{j=1}}^{\mathrm{n}}\mathrm{X}_{\mathrm{ij}}^2}$$

当 k 远小于 r 时, 称 $\mathrm{N_k}$ 为 N 的低阶近似, 其中两矩阵之差 X 的 F 范数要尽可能的小. 通过 SVD 分解后的矩阵降维的过程就是求低阶近似的过程[9].

(4) 使用降维后的矩阵构建潜在语义空间, 计算词语之前或文档之间的相似度. 本文通过向量间的夹角来判断两个对象的相似程度, 并且通过以下公式得到词与词的相似度矩阵:

$$\mathrm{R_{term}}=\bar{\mathrm{N}}\bar{\mathrm{N}}^\mathrm{T}$$

$$=(\mathrm{U_k}\sum\nolimits_\mathrm{k}\mathrm{V_k^T})(\mathrm{V_k}\sum\nolimits_\mathrm{k}^\mathrm{T}\mathrm{U_k^T})=\mathrm{U_k}\sum\nolimits_\mathrm{k}\sum\nolimits_\mathrm{k}^\mathrm{T}\mathrm{U_k^T}$$

文档与文档的相关性矩阵可以通过下式计算:

$$\mathrm{R_{term}}=\bar{\mathrm{N}}^\mathrm{T}\bar{\mathrm{N}}$$

$$=(\mathrm{U_k}\sum\nolimits_\mathrm{k}\mathrm{V_k^T})^\mathrm{T}(\mathrm{V_k}\sum\nolimits_\mathrm{k}^\mathrm{T}\mathrm{U_k^T})=\mathrm{V_k}\sum\nolimits_\mathrm{k}^\mathrm{T}\sum\nolimits_\mathrm{k}\mathrm{V_k^T}$$

4.5 语义空间的建立

语义空间建立需要收集尽可能多的与案例相关的资料, 通过不断的实验来建立良好的语义空间. 设计如下流程建立语义空间:

1) 将案例库和问答库中的文本集拆分为文档(每句话作为一个文档), 对文档进行切词、筛选、过滤停用词等操作, 统计文档中词频生成“词汇-文档”矩阵.

2) 将“词汇-文档”矩阵作加权转换, 权重计算分为局部权重、词语全局权重、文档全局权重, 考虑到权重模型组合的多样性, 将这三种权重分别定义, 并组合选择、加以对比.

3) 对加权转换后的“词汇-文档”矩阵作截断的奇异值分解, 生成潜在语义空间, 并以二进制的方式单独保存在数据库中.

4.6 总体结构

本文提出的咨询系统核心检索部分, 主要经过分词、去停用词、扩展词库、LSA 相似度计算, Lucene 信息提取等一系列步骤. 总体框架图如图 3:

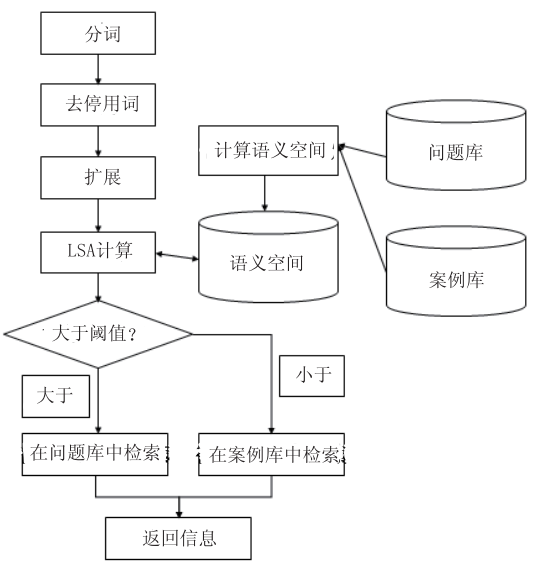

图 3 总体框图

5 实验

为了证明该算法对规模较大的问题同样具备较高的效率和有效性, 本系统使用法律出版提供的《中国指导案例》数据库, 库中包含了民事诉讼案例 10397 例, 刑事诉讼案例 2960 例, 行政诉讼案例 2136 例, 仲裁案件 368 例以及国家赔偿案例 139 例, 总共 16000 例作为系统的案例库. 问题答案库的构建通过使用实验室开发的网络爬虫, 从《法律咨询网》、《中国法律援助网》以及《法律答疑网》上面爬取信息而构建的. 问答库总共包含 328 条一对一关系的问题和答案.

构建索引的时间代价和索引所需要的空间代价是评价系统索引模块的两项重要指标.

表 1 不同规模案例库构建索引需时间和空间代价

记录数	空间代价(MB)					时间代价(S)			
	XML 文档	元组索引	关系索引	合计	单位记录(KB)	元组索引	关系索引	合计	单位记录(ms)
1000	3.01	2.16	13.9	16.06	3.29	1.7	28	29.7	5.9
3000	5.81	4.14	28	32.14	3.29	9.7	55.6	65.3	6.5
5000	11.4	7.78	55.7	63.48	3.25	55.6	107.9	163.5	8.2
10000	28.1	18.3	138	156.3	3.20	355.1	278.4	633.5	12.7
15000	56.1	36.5	286	322.5	3.30	1401.1	573.3	1974.4	19.7

从表 1 数据可以看出, 在不用数据规模下, 其单位记录的空间代价都是 3.25KB 左右, 呈线性关系; 从时间代价看, 平均检索每条数据的时间为 5.5 毫秒.

本文还分别对不同规模的案例库进行了不同关键字组合的检索. 试验结果如表 2.

表 2 不同规模案例库下不同关键字组合的检索

实验指标	关键字序号 / 规则	1	2	3	4	5	平均指标
检索单位(单位:ms)	5000	63	31	31	297	16	88
	10000	93	47	31	1031	16	244
	20000	156	47	31	3641	16	778
	50000	391	94	47	21688	31	4450
	100000	906	172	63	85344	47	17306
检索结果(单位:个)	5000	453	88	20	786	6	271
	10000	765	145	35	1694	11	530
	20000	1402	269	64	3522	23	1056
	50000	3242	643	165	9077	60	2637
	100000	6375	1271	325	18241	119	5266
单位时间(单位:ms)	5000	0.14	0.35	1.55	0.38	2.67	0.32
	10000	0.12	0.32	0.89	0.61	1.45	0.46
	20000	0.11	0.17	0.48	1.03	0.70	0.74
	50000	0.12	0.15	0.28	2.39	0.52	1.69
	100000	0.14	0.14	0.19	4.68	0.39	3.29

从表 2 中可以看出, 绝大多数检索在 20s 内都可以完成, 而平均单位检索时间基本都保持在毫秒级.

最后, 通过检索 7 组不同数量问句的答案, 考查系统的正确回答情况以及平均耗时. 试验结果如表 3.

由表 3 可知, 该系统的正确回答率保持在一个较高的水平, 可以在短时间内检索到客户需要的信息.

6 总结

答案提取是问答系统的关键之一, 本文提出的基于 LSA 和 Lucene.net 的方法, 从答案的匹配和提取两

表 3 系统正确回答情况

编号	问题数	正确答案数	正确率	平均耗时 ms
1	20	18	90.0%	122
2	26	24	92.3%	168
3	33	30	90.9%	160
4	36	34	94.4%	145
5	40	37	92.5%	137
6	22	21	95.5%	165
7	19	17	89.5%	186
平均			92.2%	153.14

方面做了深入的改进. 基于 Lucene.Net 的搜索方案可以高效快速的建立索引, 这种索引不仅仅局限于数据库, 也包含了我们日常工作中经常遇到的文件格式. 下一步工作是结合分布式框架 Hadoop, 建立分布式数据检索系统. 基于潜在语义分析的问题和答案句子相似度计算方法, 对词的同义和多义现象有较好的处理效果, 答案提取实验结果也说明了这一点, 但由于 LSA 的基础是基于词频统计方法, 没有考虑词在答案句子中的位置分布对答案句子的影响, 而且 LSA 方法本身也存在计算量大、所需存储空间大等缺点, 因此该方法还存在一定的局限性. 进一步的研究将结合词的位置分布和词的语义关系进行答案句子和答案实体的抽取.

参考文献

1 余正涛,樊孝忠.基于潜在语义分析的汉语问答系统答案提取.计算机学报,2006,29(10):1889−1893.

2 刘江平,薛河儒. 基于 Web 的智能答疑系统的设计和实现. 网络与通信,2012,28(3):121−123.

3 李园伟,宁可为,王炜.分布式自动答疑系统.计算机系统应用, 2012,21(7):22−29.

4 孔维亭,闫宏印.基于 Lucene 的自动答疑系统的设计.电脑开发与应用,2012,25(4):32−36.

5 吴秀梅.基于潜在语义分析和最大墒的中文情感分析研究[学位论文]. 北京:北京交通大学,2011.

6 王丛林.在线自动答疑系统设计与开发的研究[学位论文]. 长春:东北师范大学,2010.

7 Ng.A. Indexing and searching image files: using Lucene.NET long with open-source libraries. Dr. Dobb's Journal, 2008, 33(10): 52−55.

8 Zhang CH. Research and implementation of full-text etrieval system using compass based on Lucene.Advances in Intelligent Systems and Computing, 2013, 181: 349−356.

9 Su TY. Chinese full-text retrieval system based on Lucene. Computer Engineering, 2007, 33(23): 94−96.

10 Yin PP. Evaluation of literature frontier based on latent semantic analysis. Proc. the 2012 IEEE Symposium on Robotics and Applications (ISRA). 2012. 403−406.

11 顾益军,樊孝忠.中文停用词表的自动选取.北京理工大学学报,2005,25(4):337−341.

(上接第 96 页)

参考文献

1 张世明,杨寅春.基于角色的访问控制技术在大型系统中的应用.计算机工程与设计,2006,27(19):3723−3725.

2 陈辉,赵洪升,张艳春. Struts+Spring+ Hibernate 框架的整合实现.河南大学学报(自然科学版),2010,40(6):642-645.

3 曹渠江,陈真. Struts2 框架整合 Spring 框架在文件上传下载中的应用.上海理工大学学报,2009,31(2):169−172.

4 孙更新,宾晟,宫生文.Java 程序开发大全.北京:中国铁道出版社,2010:214−215.

5 闫宏印,张卫争,刘超慧.开源框架下 Web 应用分层的设计与实现.计算机工程与设计,2008,29(23):6023−6028.

6 陈翠娥.Java 单例模式应用研究.长沙民政职业技术学院学报,2010,17(1):114−116.

7 刘艳春,洪晓慧.Struts2 框架核心配置文件的研究与应用. 计算机技术与发展,2013,23(2):77−81.

8 王恒娜.访问局部性原理在 Cache 存储系统中的作用.安徽大学学报(自然科学版),2005,29(1): 27−3