一种基于语义匹配的 Web 信息提取方法研究

张茂元[1,2]　邹春燕[3]　卢正鼎[1]
[1](华中科技大学计算机科学与技术学院,武汉 430074)
[2](华中科技大学管理学院,武汉 430074)
[3](华中师范大学外国语学院,武汉 430079)
E-mail:zmydragon@163.com

摘　要　为了较好地解决信息过量难以消化、汉语词的歧义划分、Web 信息形式不一致并且难以辨识的问题,文章提出了一种基于语义匹配的 Web 信息提取方法。该方法融合了网页分类、汉语分词、语义信息匹配方法,并给出了一种义素相似度,进而提出了一种基于语义的信息匹配方法来识别和提取网页信息项。基于这种 Web 信息提取方法的网上药品信息监管系统 Web-MIND 能够提取出网上药品广告的信息项,并具有较高的准确率。

关键词　信息提取　语义　匹配

文章编号 1002-8331-(2006)23-0141-03　**文献标识码** A　**中图分类号** TP391

An Information Extraction Based on Semantic Matching for Web Pages

Zhang Maoyuan[1,2]　Zou Chunyan[3]　Lu Zhengding[1]
[1](Department of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074)
[2](School of Management, Huazhong University of Science and Technology, Wuhan 430074)
[3](School of Foreign Language, Huazhong Normal University, Wuhan 430079)

Abstract: Some problems exist in all these Web information, for example: difficulty in processing excessive information, Chinese word segmentation for the ambiguous words, the information of variable formats, and the recognition of information. In order to solve those problems, an information extraction of web pages based on semantic matching is proposed in this paper. The extraction method integrates the classification method of Web pages, segmentation method of Chinese words and semantic-matching method of information. Moreover, the extraction method proposes a sememe based similarity and then puts forward a novel semantic matching method of information, which is used to recognize and extract the information items of Web pages. Based on the extraction method, the monitor system for Web information of drugs can extract the information items of drug advertisement in Web with high accuracy.

Keywords: information extraction, semantic, matching

1　引言

面对 Web 信息的飞速增长，人们迫切感到需要新的技术和工具以便从 Web 数据源中智能地、自动地抽取有价值的知识信息。如何快速、准确地获得有价值的网络信息,如何从这些海量数据中发现知识,这就要求有一个高效、高准确率的 Web 信息提取工具[1]。

Internet 在飞速地发展的同时,不仅使人们获得大量信息,也给 Web 信息提取带来了一些问题:

(1) 信息过量难以消化。Web 信息以网页的形式存在,而 Web 上网页的数量超过 3 亿。因此 Web 信息提取需要对网页进行分类,过滤掉不需要的网页,从而缩小要处理的网页集合。

(2) 汉语词的歧义划分。Web 信息有西文信息,也有汉语信息。西文的词与词之间有分隔符,但汉语的词与词之间没有分隔符。由于词在一定程度上可以体现信息的语义,因此汉语分词的正确率在一定程度上可以影响 Web 信息提取的正确率。

(3) Web 信息形式不一致,并且难以辨识。各个网站在发布网页信息时,不采用统一的网页风格,而采用各自喜好的网页风格来制作网页,并且有的网站会不定期地更换自己的网页风格。因此对变化的网页信息结构,需要研究它的适应能力和信息项的匹配方法,来提高 Web 信息提取的准确率。

Web 信息提取中词匹配的方法可分为基于词频的方法和基于词分类关系的方法。基于词频的词匹配方法[2~4],用词在文档中出现的共同程度来体现词间的相似度。这类方法体现词在出现分布上的语义关系,但未考虑词分类学中的结构关系。基于词分类关系的方法是建立在词汇语义网络中的分类关系层次上。一种基于实体论的词相似法[5]和一种基于语义的模糊匹

基金项目:国家自然科学基金资助项目(编号:60403027)
作者简介:张茂元(1975-),博士后,讲师,主要研究方向为信息检索与提取、自然语言处理和信息管理。邹春燕(1978-),硕士,讲师,主要研究方向为语义学。卢正鼎(1944-),教授,博导,主要研究方向为信息系统。

配方法[6],建立在词汇语义网络中层次关系间的距离因素,但未考虑层次关系中的深度因素。基于不同实体论的词相似法[7]考虑了层次关系中的深度因素,并获得了较好的效果,但面对新出现的词,词汇语义网络就需要扩充。

目前已有的 Web 信息提取方法[8~11]主要致力解决 Web 信息形式不一致的问题,为了较好地解决信息过量难以消化、汉语词的歧义划分、Web 信息形式不一致并且难以辨识的问题,文中第 2 节提出了一种基于语义匹配的 Web 信息提取方法,该方法融合了网页分类、汉语分词、语义信息匹配方法。针对语义信息匹配,文中第 3 节给出一种义素相似度,并在此基础上提出一种基于语义的信息匹配方法。文中第 4 节给出了基于语义匹配的 Web 信息提取方法的实验,从实验结果上看,该方法具有较高的准确率。

2 基于语义匹配的 Web 信息提取

2.1 系统结构

如图 1 所示,基于语义匹配的 Web 信息提取模型由模糊网页分类、基于学习的网页信息提取、语境汉语分词、语义信息匹配和分布式主动数据库五个部分组成。其输入是搜索到的网页,输出是要得到的网页中的信息项。

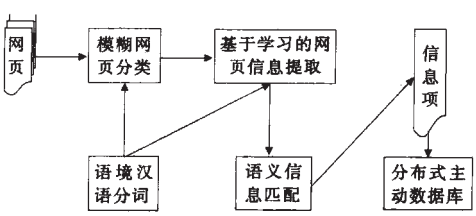

图 1 Web 信息提取的系统结构

模糊网页分类模块对网页进行模糊分类,初步过滤掉不相关的网页,缩小处理集合;基于学习的网页信息提取模块,分析网页标记,用树型结构(Document Object Model)表示 HTML 网页的布局,并用该模型来学习和识别网页结构模式,从而依据模型来提取 HTML 网页中的丰富数据部分;语境汉语分词模块为模糊网页分类模块、基于学习的网页信息提取模块进行汉语分词预处理;语义信息匹配模块对提取到的信息进行语义匹配,识别信息中各个信息项所属的信息项目类别,并提取出信息项;提取出的 Web 信息项存入到分布式主动数据库中,由数据库主动地处理这条信息,如比较历史信息来检测信息的一致性、合法性等处理,这就使系统能够为信息预测、信息预警等后期处理提供一定的及时性。

2.2 信息监管系统 Web-MIND

在这种提取模型基础上,自主研发了网上药品信息监管系统 Web-MIND。该系统能够完成 Internet 上广告信息的查找、过滤、提取、违法内容的审定。该系统用作者已研究出的一种基于特征选取及模糊学习的网页分类方法[12]对网页进行分类,过滤掉非药品广告的网页,并用作者已研究出的一种基于语境的汉语分词方法[13]进行汉语分词预处理。然后系统采用 DSE(Data-rich Section Extraction)算法[14]提取出药品网页信息。接着系统用下一节提出的基于语义的信息匹配方法对提取到的药品广告信息进行语义匹配,识别信息中各个信息项所属的信息项目类别,并按所属的项目类别提取出各信息项,然后存入到数据库。最后系统把提取到的药品广告信息与国药局的法规,用下一节提出的基于语义的信息匹配方法对广告信息项和法规进行语义匹配,来检测药品广告信息的合法性,并把药品广告信息和检测结果存入到数据库中。当药品广告信息一旦存入主动数据库时,数据库就能立即、主动地把信息与它的历史信息进行分析,这样得到的分析结果对网上药品广告信息监管的预防和预警是有所帮助的。在分布式主动数据库模块方面,作者已从理论上研究了一种面向 Agent 的分布式主动数据库框架[15]。

3 基于语义的信息匹配方法

3.1 义素网络

《知网》(HowNet)是一个网状的有机知识系统,以汉语和英语的词语所代表的概念为描述对象,来表示概念与概念以及概念属性之间的关系。在知网中,“义素”是从所有汉语词汇中提炼出的可以用来描述其它词汇的不可再分的基本元素,每一个概念是通过一组义素来表示的。

从药品信息的特征词(如功能、治疗等)中,提取出每个义素,组成药品信息义素集合。然后按照 HowNet 的构建原理,对义素集合构建药品信息的语义网络 HowNet—medicine。

3.2 义素相似度函数

3.2.1 基于语义路径的相似度

由于语义网络的构建应用了词汇分类法,义素间的语义距离可以用义素间连接边的数量来表示,所以义素相似度可以用语义路径长度来计算。

定义 1 设两个义素 $seme_1$ 和 $seme_2$ 之间的路径长度为 L,基于语义路径的相似度为:

$$Sim_1(seme_1, seme_2)=f_1(l)=e^{-\alpha l} \quad (1)$$

其中 $\alpha>0$ 是常数,$l\in[0,+\infty)$。

3.2.2 改进的语义相似度

尽管基于语义路径的相似度在一些问题上取得了较好的结果,但在大型或通用的语义网络应用中,这种计算方法在准确度上存在一定的误差。为了改进这个不足,相似度计算还需引入更多的语义网络结构信息。从直观角度上,位于语义网络中较高层次的义素含有较通用的语义和较弱的相似度,而位于语义网络中较低层次的义素含有较具体的语义和较强的相似度。所以,在计算相似度时,义素的层次深度应当得到考虑。

定义 2 $f_2(h_1,h_2)=\dfrac{e^{\frac{\beta(h_1+h_2)}{2}}-e^{\frac{-\beta(h_1+h_2)}{2}}}{e^{\frac{\beta(h_1+h_2)}{2}}+e^{\frac{-\beta(h_1+h_2)}{2}}}$,其中 $\beta>0$ 是常数,h_1,$h_2\in[0,+\infty)$。

定义 3 设两个义素 $seme_1$ 和 $seme_2$ 之间的路径长度为 L,且它们的层次深度分别为 h_1 和 h_2,则改进的义素相似度为:

$$Sim_2(seme_1,seme_2)=f(f_1(l),f_2(h_1,h_2))=f_1(l)\times f_2(h_1,h_2) \quad (2)$$

3.3 基于义素的词相似度

定义 4 设词 w 含有 n 个义素 $seme_1,seme_2,\cdots,seme_n$,则该词可以用义素向量 $semeV=(seme_1,seme_2,\cdots,seme_n)$ 表示。

定义 5 设义素 seme 和义素向量 $semeV=(seme_1,seme_2,\cdots,seme_n)$,则义素与义素向量之间相似度是 $Sim_3(seme,semeV)=\max_{j=1}^{n} Sim_2(seme,seme_j)$。

定义 6 设词 w_1 的义素向量是 $semeV_1=(seme_{11},seme_{12},\cdots,seme_{1n})$,词 w_2 的义素向量是 $semeV_2=(seme_{21},seme_{22},\cdots,seme_{2n})$,则基于义素的词 w_1 与词 w_2 的相似度是:

$$Sim_4(w_1,w_2)=\frac{1}{|semeV_1|}\sum_{i=1}^{m} Sim_3(seme_i,semeV_2) \quad (3)$$

其中，义素向量 $semeV_1$ 的模 $|semeV_1|=m_1$。

人类知识发展中出现的新词，像旧词一样含有义素，因此新词可用义素向量来表示。这样，新词和旧词之间的相似度可用式(3)来计算，所以基于义素的词相似度不仅用于旧词，同样能适用于新词，因而具有较强的自适应性。

3.4 基于语义的信息匹配

如图 2 所示，基于语义的信息匹配模块分为 4 层。第一层是义素转换层，用义素向量 $semeV_i$ 表示从各个信息项 $Item_i$ 中提取的信息关键词 w_i；第二层用式(3)计算义素向量 $semeV_i$ 与信息项类特征词 T_j 的相似度 m_{ij}；第三层采用竞争机制，比较第二层的结果，如果 m_{ij} 最大，则使 $y_j=i$，但如果所有值都很小，则使 $y_j=0$；第四层合并各分量，输出信息项匹配矢量。

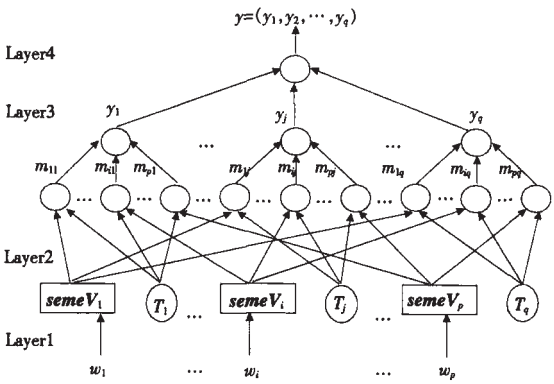

图 2 基于语义的信息匹配系统结构图

4 实验

用于网上药品信息监管系统 Web-MIND 测试的数据集来自当代中医网、大众健康网、三九健康网、37c 医学网、河南海虹医药电子商务网、成都医药资源网、京卫医药网、吉林海虹医药电子商务网 8 个医药网站网页。Web-MIND 系统分别采用基于词语义网络距离的语义匹配方法和基于义素的语义匹配方法(α=0.2, β=0.6)来提取出药品广告的信息项。这两种方法分别定义为方法 A 和 B。实验用到 1 060 张药品广告信息网页，提取结果如表 1：

表 1 两种方法的信息项提取结果

药品广告信息项的类别	数量	方法 A		方法 B	
		正确数	准确率	正确数	准确率
药名信息	1 060	896	84.5%	957	90.3%
功能信息	1 060	882	83.2%	951	89.7%
厂商信息	730	618	84.7%	658	90.1%
禁忌信息	820	676	82.4%	731	89.1%
总计	3 670	3 072	83.7%	3 297	89.8%

从平均准确率角度上看，方法 B 的平均准确率是 89.8%，比方法 A 高 6%。这表明义素相似度，在一定程度上提高了信息项提取的准确率。从最大准确率与最小准确率之差角度上看，方法 A 的最大准确率与最小准确率之差是 2.3%，而方法 B 的差值是 1.2%。因此方法 B 比方法 A 在准确率上具有更好的稳定性，这表明义素相似度，在一定程度上也提高了信息项提取结果的稳定性。

综上，基于义素语义匹配的 Web 信息提取方法不仅有较高的准确率，而且有较小的最大准确率与最小准确率之差。因此，它是一种较好的 Web 信息提取方法。

5 结束语

这里提出的一种基于语义匹配的 Web 信息提取方法，融合了网页分类、汉语分词、基于学习的 Web 信息提取等方法，并给出了一种基于义素的信息语义匹配方法。在这种提取方法基础上，自主研发了网上药品信息监管系统 Web-MIND，该原型系统能够完成 Internet 上广告信息的查找、过滤、提取以及违法内容的审定。(收稿日期：2005 年 12 月)

参考文献

1.李永平，张茂元.基于并行模糊归类的网页信息提取方法研究[J].计算机工程与应用，2003；39(21)：23~24

2.Shen H T，Shu Y，Yu B.Efficient semantic-based content search in P2P network[J].IEEE Transactions on Knowledge and Data Engineering，2004；16(7)：813~826

3.Yi Shanzhen，Huang Bo，Tatchan Weng.XML application schema matching using similarity measure and relaxation labeling[J].Information Sciences，2005；169(1-2)：27~46

4.Nakashima T.Classification of characteristic words of electronic newspaper based on the directed relation[C].In：2001 IEEE Pacific Rim Conference on Communications，Computers and Signal Processing，Victoria，B C，Canada：IEEE Computer Society Press，2001：591~594

5.Vladimir A O.Ontology based semantic similarity comparison of documents[C].In：14th International Workshop on Database and Expert Systems Applications(DEXA'03)，Prague，Czech Republic：IEEE Computer Society，2003：735~738

6.程莉，卢正鼎，文坤梅.基于语义的模糊匹配探索与应用[J].华中科技大学学报(自然科学版)，2003；31(2)：23~25

7.Rodriguez M A，Egenhofer M J.Determining semantic similarity among entity classes from different ontologies[J].IEEE Transactions on Knowledge and Data Engineering，2003；15(2)：442~456

8.Gao X，Zhang M，Andreae P.Learning information extraction patterns from tabular Web pages without manual labeling[C].In：Jiming Liu ed.2003 IEEE/WIC International Conference on Web Intelligence，Halifax，Canada：IEEE Computer Society，2003：495~498

9.Alani H，Sanghee Kim，Millard D E.Automatic ontology-based knowledge extraction from Web documents[J].IEEE Intelligent Systems，2003；18(1)：14~21

10.Manjula D，Aghila G，Geetha T V.Document knowledge representation using description logics for information extraction and querying[C].In：2003 International Conference on Information Technology：Coding and Computing[Computers and Communications]，The Orleans，Las Vegas，Nevada：IEEE Computer Society，2003：189~193

11.Kao Hungyu，Lin Shianhua，Ho Janming et al.Mining Web informative structures and contents based on entropy analysis[J].IEEE Transactions on Knowledge and Data Engineering，2004；16(1)：41~55

12.张茂元，卢正鼎.基于特征选取及模糊学习的网页分类方法研究[J].小型微型计算机系统，2004；25(7)：1397~1400

13.Zhang Mao-yuan，Lu Zheng-ding，Zou Chun-yan.A Chinese Word Segmentation Based on Language Situation in Processing Ambiguous Words[J].Information Sciences，Elsevier，2004；162(3-4)：275~285

14.Jiying Wang，Lochovsky F H.Data-rich section extraction from HTML pages[C].In：Wee Keong，Tok-Wang Ling eds.Third International Conference on Web Information Systems Engineering(Workshops)，Singapore：IEEE Computer Society，2002：313~322

15.张茂元，卢正鼎.一种 Agent 数据库系统框架及其规则并行算法[J].软件学报，2004；15(8)：1157~1164