文章编号: 1000-6788(2010)01-0096-09 中图分类号: TP18 文献标志码: A

AIS — 基于文本挖掘的增强型 Web 信息处理技术

张 文[1,2], 唐锡晋[3], 吉田武稔[2]

(1. 中国科学院 软件研究所互联网实验室, 北京 100190; 2. 北陆先端科学技术大学院大学, 日本石川县 923-1211;
3. 中国科学院 数学与系统科学研究院, 北京 100190)

摘 要 回顾了中文和英文语言环境下的 Web 文本挖掘现状, 阐明了其现阶段的特点和技术瓶颈. 之后提出了一种基于 Web 文本挖掘的网页内容挖掘技术: AIS (Augmented information support), 介绍了相关实现所涉及的基础技术和功能. 最后将 AIS 技术应用于香山科学会议网站, 开发了 AIS4XSSC 文本挖掘系统并展示了现阶段其主要功能. 实践表明 AIS 技术能够从大量的 Web 文本中有效提炼信息, 提高用户检索效率并向用户推送有价值的信息.

关键词 Web 文本挖掘; 知识发现; AIS; 综合集成研讨厅; 香山科学会议

AIS: An approach to Web information processing based on Web text mining

ZHANG Wen[1,2], TANG Xi-jin[3], YOSHIDA Taketoshi[2]

(1. Lab for Internet Software Technologies, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China;
2. Japan Advanced Institute of Science and Technology, Ishikawa 923-1211, Japan;
3. Academy of Mathematics and of Systems Science, Chinese Academy of Sciences, Beijing 100190, China)

Abstract Web text mining (WTM) is a technology for information support as one component of the machine system of HWMSE. Concerning the deficiencies of current search engine for retrieval of WWW, improvements are expected. In this paper, a brief review on recent WTM developments was presented at first. Then a technology on augmented information support, AIS, was proposed to cope with "information explosion" based on WTM technologies. Finally, AIS is applied to the development of the AIS4XSSC (AIS for Xiangshan Science Conference) system, which is customized for information retrieval and knowledge discovery from XSSC Website. The practical application demonstrates that AIS is useful to extract information from Web documents and improve the performance of information retrieval.

Keywords Web text mining; knowledge discovery; AIS; HWMSE; Xiangshan science conference

1 介绍

随着计算机和互联网逐渐成为科学研究和经济发展不可或缺的工具,Web 文本日益膨胀, 以致 "信息爆炸". 据统计, 截至到 2007 年 2 月, 全球共有超过 7 千万的互联网站点. 这些网站拥有总共约 297 亿网页, 其中约 80%是文本信息[1−2]. 数量庞大的 Web 文本资料中蕴含了大量的潜在有价值的知识, 但是, 如何从中快速、准确、有效地发现对用户有用的知识是摆在研究人员面前的一个重要课题. Web 文本挖掘可以帮助用户检索和浏览所需要的信息, 并能够发现 Web 文本的模式和知识.

综合集成研讨厅 (Hall for workshop of metasynthetic engineering, HWMSE)[3] 是思维科学的一项具体的应用技术, 它将专家体系、机器体系和数据体系有机结合起来组成智能系统. 如图 1 所示, Web 文本挖掘为 HWMSE 的成功应用提供了基础支撑技术. 首先, 它能够为研讨厅专家体系提供来自于 Web 的信息支

收稿日期: 2008-11-07
资助项目: 国家自然科学基金 (70571078)
作者简介: 张文 (1981–), 通讯作者, 男, 博士、助理研究员; 唐锡晋 (1967–), 女, 博士, 研究员.

持, 并对这些信息进行加工整理, 使专家方便及时的掌握所需要的研讨信息; 其次, 对于研讨厅中的知识体系, Web 文本挖掘能够帮助专家从 Internet 网络中发现新的知识和模式, 及时补充更新知识体系中的知识; 最后, 对于机器体系, Web 文本挖掘能够自动从网络中收集、整理信息, 构建机器体系的基础数据平台, 丰富研讨厅机器体系. 反过来, 研讨厅专家集体智慧的涌现又能够从应用的层面帮助提高 Web 文本挖掘技术, 尤其是在 Web 文本挖掘知识模式的解释方面. Web 文本挖掘属于信息技术, 而综合集成属于思维科学的范畴, 将这二者的交叉融合必定能够产生出新的内容, 实现 $1 + 1 > 2$ 的效果. 本文所阐述的 AIS 技术正是在 HWMSE 体系思想的指导下所开发出来的.

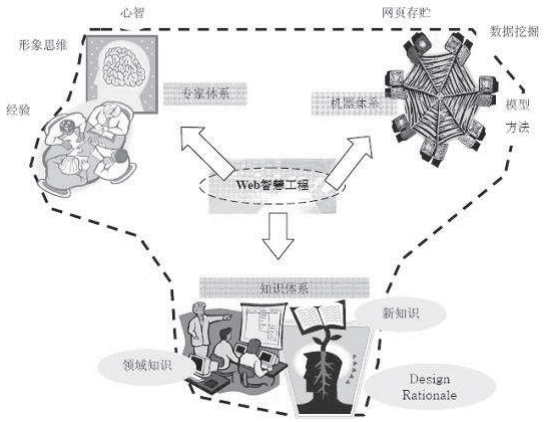

图 1　综合集成 Web 智慧工程, Web 文本挖掘丰富了综合集成研讨厅的机器体系 [4]

目前可用的 Web 搜索引擎部分解决了互联网中资源定位和查找的问题, 但其功能还有待从 3 个方面提高: 时下应用最为广泛的搜索引擎一般仅基于文字字符匹配, 而不是基于语义概念的搜索; 传统的搜索引擎的返回结果信息需要提炼; 传统的搜索引擎只能列出返回结果, 而不能给出搜索结果之间的关联关系.

基于以上三点原因, 需要开发和利用比现有搜索引擎更加智能的技术, 从大量的 Web 文本中发现有价值和可理解的模式和知识. 这种技术就是知识发现 (Knowledge discovery). 数据挖掘 (Data mining 或者 Knowledge discovery from database) 面向的对象是结构化的关系型数据库, 它并不适合 Web 文档半结构化和非结构化的特征. 因此, 我们只能采用文本挖掘 (Knowledge discovery from texts), 根据用户的需求, 从大量的 Web 文本中快速, 有效的发现有用的知识.

需要指出的是, 虽然 Web 挖掘有多方面的任务, 如使用日志挖掘, 结构挖掘等. 但针对 Web 中广泛存在的半结构化和非结构化的文本, 本文的关注点在于 Web 文本挖掘, 即从大量 Web 纯文本中发现出某种有用的知识和模式. 本文并没有涉及到挖掘 Web 中图片, 声音, 链接等其他信息.

由于文本挖掘是一个新兴的研究领域, 不同背景的研究人员由于其研究出发点不同而对于文本挖掘的定义也不尽相同, 主要有以下三种定义 [5]:

1. 文本挖掘就是信息抽取 (Information extraction). 这种定义将文本挖掘视作为按照某种预先定义的模板 (Predefined template) 从文本中抽取用户关注的信息.

2. 文本挖掘就是对文本进行数据挖掘. 这种定义将文本挖掘视作数据挖掘的扩展应用. 也就是, 应用统计机器学习的方法在文本中发现有用的模式. 为了达成此目的, 需要对文本进行一系列的预处理, 例如信息抽取, 自然语言处理 (Natural language processing) 或者其他方法等, 从文本中抽取结构化的数据从而应用数据挖掘的方法进行模式识别.

3. 文本挖掘就是在文本中进行知识发现. 这种定义强调从大量的文本集合中抽取和发现面向用户需求的知识. 它包括两个方面, 其一是文本数据挖掘, 其二是语学工程 (Language engineering). 由于各种语言和文本体裁的差异, 在实际的文本挖掘过程中, 必须考虑各种语言模型和融合语料库学习的方法.

本文所采用的定义为以上的第三种, 即 Web 文本挖掘就是利用数据挖掘, 信息检索, 语学工程等技术在 Web 文本中进行知识发现.

2　先行研究

2.1　英文 Web 文本挖掘研究现状简介

在英文文本挖掘方面, Hearst[6] 首先提出了将数据挖掘技术引入到以文本挖掘, 讨论了数据挖掘, 信息检索和基于语料库的计算语言技术之间的关系, 并阐述了文本挖掘技术在疾病治疗, 专利管理, 基因工程等方面潜在的应用价值. Feldman[7] 等利用 Reuters-21578 文本集合中的文本关键词分布以及关键词之间的同现 (Co-occurrence) 来进行知识发现, 他们的方法为利用数据挖掘的方法在文本中进行知识发现提供了一种技术思路. Hotho[8] 等提出了利用的策略来利用 Ontology 改善文本表示向量, 从而提高文本聚类的效果. 在

应用方面, Yang 和 Lee[9] 利用 SOM(Self-organizing mapping) 技术来进行文本和关键词进行聚类, 利用文本挖掘技术自动构建文本之间的超级链接. Lo[10] 利用 SVM (Support vector machine) 来进行顾客抱怨信息自动分类, 以提高客服的工作效率和客户的服务满意度. Zhang 和 Nasraoui[11] 利用搜索引擎的查询日志来动态获取用户的查询关注点和搜索习惯, 从而提出了查询推荐 (Query recommendation) 和查询提炼 (Query refinement) 方法. Liu 等 [12] 利用 Web 文本挖掘技术开发了一个网络金融信息自动采集和处理系统, 用以进行观测网络上关于某类金融产品的评价, 从而预测未来的金融行情.

相比于中文 Web 文本挖掘, 英文 Web 文本挖掘由于其基础 Web 技术和自然语言处理技术得到了广泛的研究, 并开发了一些便于高级应用的工具如 WordNet, 加之公开的 Web 文档语料库的完善. 因此, 针对英文 Web 文本挖掘提出的算法、模型、应用技术等可以迅速得到验证. 所以, 英文 Web 文本挖掘的研究一般采取算法, 实验, 评价的研究思路.

2.2 中文 Web 文本挖掘的研究现状简介

自从文本挖掘被提出以来, 其研究和发展一直受到中文信息处理学界的关注. 在文本挖掘的理论方面, 文 [13] 将概念词典 WordNet 应用于文本聚类, 分析了中文文本特征提取和中文文本摘要并在此基础上提出了一种面向中文文本挖掘的模型. 文 [14] 提出了一种 Web 使用挖掘的框架, 提出了根据用户访问 Web 的规律来改进网站的设计从而推出个性化的用户服务, 分析了网站结构和内容对 Web 使用挖掘的影响, 并对 Web 使用挖掘在电子商务中可能的应用进行了详细论述. 文 [15] 对 Internet 上的文本进行了综述性的介绍, 给出了 Internet 上文本挖掘的一般处理过程, 并对每个处理过程中所使用的技术进行了详细的讨论. 文 [16] 回顾了文本分类技术的研究状况, 提出了结合网页的结构信息选择合理的网页表示方式和分类算法.

在文本挖掘的应用方面, 文 [17] 介绍了一个 Web 文本挖掘系统的原型 WebMiner, 该系统采用了多 agent 体系结构, 将多维文本分析与文本挖掘这两种技术有机地结合起来, 以帮助用户快速、有效地挖掘 Web 上的 HTML 文档. 文 [18] 介绍一个网络挖掘原型系统 WebME, 包括其系统结构、主要功能和特点, 并提出了进一步完善的一些设想. 文 [19] 开发了一个专门针对 Web 上有关城市信息的信息检索系统, 并利用同义词典和查询扩展技术改善了中文网页查找的效果.

总体上, 中文 Web 文本挖掘研究方面有关中文文本挖掘的理论方法提出很多. 由于中文处理难度较大, 特别是基础公共的中文自然语言处理工具和平台 (例如汉语词法分析工具和汉语概念词典) 的缺乏, 这些理论并未得到很好的验证.

3 AIS 技术

3.1 AIS 的文本挖掘过程

本文探讨的 AIS 技术中采用的 Web 文本挖掘处理过程如图 2 所示. 图 2 中描述的过程与文 [3–5] 中描述的 Internet 上的文本挖掘处理过程类似, 其主要区别在于本文的过程增加了 Web 文本收集的功能. 这是因为, 作者认为 Web 文本挖掘过程应该是一个实时, 动态的过程, 所以需要实时从 Web 上收集、分析、整理数据, 进行 Web 文本知识发现.

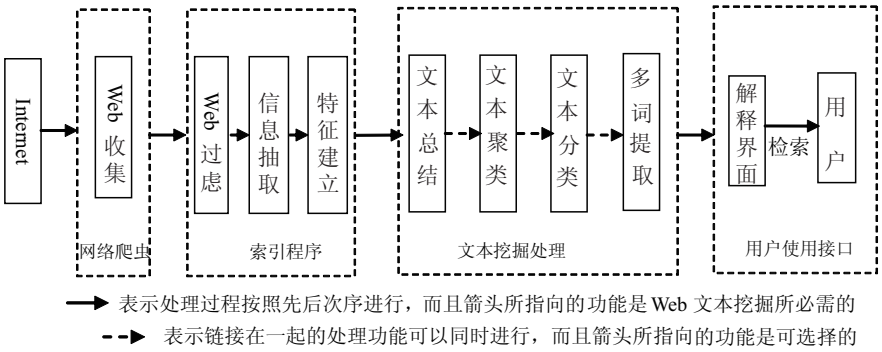

图 2 AIS 的 Web 文本挖掘处理过程

3.2 应用于 AIS 中的文本挖掘处理技术

虽然图 2 中所描述的各个处理功能在针对具体 Web 站点应用时在处理细节上会有所差别, 但是在总体实现上却大致类似, 现做如下说明.

网络爬虫的主要功能是根据给定的种子站点和深度, 按照 Web 页面之间的 URL 链接来下载 Web 页面. 经典的 Web 爬虫设计程序如算法 1 所示.

索引程序功能包括三个部分: 过滤用户不感兴趣的 Web 页面; 选择关键词索引 Web 页面; 根据 Web 页面结构从 Web 文本进行信息抽取 (Information extraction).

文本挖掘处理根据用户的需求选择合适的文本挖掘技术来从 Web 文本集合中发现知识. 这里所说的文本挖掘技术既包括了针对结构化数据的数据挖掘技术, 例如聚类和分类, 也包括了专门针对文本的自然语言处理技术, 例如文本总结和语义分析.

用户使用接口的主要功能是将 Web 文本挖掘的结果进行有效组织, 以用户可以理解的方式呈现出来, 并提供给用户快捷、方便、稳定的检索和查询界面.

输入:

I, 初始化爬虫的种子站点;

D, 深度, 也就是爬虫将会下载的页面到种子站点之间的距离;

输出:

F, 爬虫从下载的页面中解析出来的 URL 地址队列;

处理过程:

```
Begin
    For each URL i in I
        Enqueue(i,F); //将初始地址添加到 F 中;
    End
    j=0;
    While F is not empty and j < D
        u = Dequeue(F); //从 F 队列中弹出一个 URL 地址;
        if u has not been processed
            Get(u); //下载 u 所表示的 URL 地址下的 Web 页面;
            Extract the hyperlinks;
            Let U be the set of hyperlinks extracted;
            For each u in U
                Enqueue(u,F); //将解析出来的地址保存到 F 中;
            End
        End
    End
End
```

算法1　经典的Web 爬虫算法程序

4 AIS 在香山科学会议网站上的应用 —AIS4XSSC 系统

为了实现本文阐述的 AIS 技术, 作者自 2006 年起针对香山科学会议网站 (http://www.xssc.ac.cn/) 开发了 AIS4XSSC(Augmented information support for Xiangshan science conference)Web 文本挖掘系统.

4.1 AIS4XSSC 系统架构

AIS4XSSC 系统架构如图 3 所示. 该系统采用 B/S 架构, 首先将 Web 上的数据收集到本地, 然后进行信息抽取和挖掘, 最后通过 Web 发布文本挖掘的结果. 该系统包括三个组件: Web 站点信息采集, 本地存储处理和 Web 集成服务. Web 站点信息采集采用了网络爬虫技术, 从香山科学会议网站上进行诚勉下载. 本地存储处理用来索引下载的 Web 页面, 信息抽取和 Web 文本挖掘. Web 集成服务用于推送系统挖掘出的模式与知识.

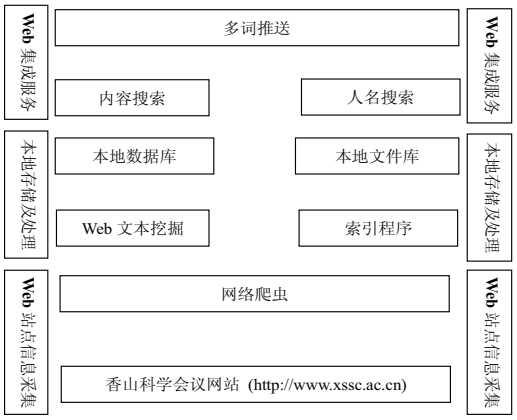

图 3　AIS4XSSC 文本挖掘系统

与传统的搜索引擎不同的是,AIS4XSSC 采用了 Web 挖掘技术对 Web 网页信息进行了更为深入的处理. 为了能够让用户快速浏览网页信息, AIS4XSSC 对各个网页进行了自动文摘, 概括网页内容. 目的是对香山科学会议的历史记录进行分门别类, 合理组织和再现. 为了让用户能够重点关注香山科学会议的参会人员信息, AIS4XSSC 利用信息抽取技术从网页中抽取了关于与会人员的信息; 开发了关于人名信息的推送 —— 当用户输入的关键词中含有某个曾经在香山科学会议参加会议的人的名字的时候, 系统会自动推送出关于这个人在香山科学会议的参会情况; 其目的是能够让用户对某位香山科学会议的专家进行参会历史追踪. 为了让用户了解与某次会议主题相关的其它会议, AIS4XSSC 利用文本聚类技术将相关主题的会议信息聚在一起, 让用户把握某类主题会议在香山科学会议上的进展情况. 其目的是试图利用现有的文本挖掘技术从香山科学会议的历史记录中帮助用户发现或者提炼某种模式、规律或者见解. 为了让用户了解香山科学会议上的专业术语, 我们开发了多词 (Multi-word expression) 推送技术. 当用户输入的关键词包含某个科技专门术语的一

部分时, 系统会自动推送出香山科学会议中含有这个关键词的所有技术术语.

4.2 AIS4XSSC 文本挖掘技术分解

4.2.1 从香山科学会议网站上进行 Web 收集

利用 2.2 节中描述的经典爬虫算法, 设计了网络爬虫, 并将香山科学会议网站的首页设为种子站点, 深度为 10, 进行 Web 页面收集. 将网络爬虫深度设置为 10 的原因是本文仅关注香山科学会议的网页, 而不关心香山科学会议以外的网站的页面, 因此, 有必要尽量保证爬虫运行在香山科学会议网站之内. 通过以上设置, 网络爬虫总共从香山科学会议的网站上下载了 646 个 Web 页面, 经过 Web 过滤, 保留了其中的 208 个页面作为进一步处理的 Web 文本集合.

4.2.2 从香山科学会议网页中进行信息抽取

图 4 是香山科学会议页面的基本结构. 从中可以看出, 该页面主要由会议标题, 会议内容和与会人员三部分组成. 基于此模板结构, 这三部分的信息以及 HTML 标签可以被分别的从页面中抽取出来, 如图 4–7 所示. 从 HTML 页面中进行信息抽取首先须将 HTML 标签和页面的文本内容进行分离. 虽然现有的 DOM 方法能够较好的从 Web 文本中进行信息抽取[20], 但由于页面中存在错误的 HTML 标签和 HTML 本省存在不符合 DOM 树规则的标签, 如 “< li >”,“< br >” 等, 使得此种方法并不能够达完全正确的区分标签和内容. 为此, 本文采用文字匹配规则来匹配标签 “<>” 之间的内容和之外的内容, 从而区分 HTML 标签和页面的内容.

图 4 香山科学会议 Web 页面的结构

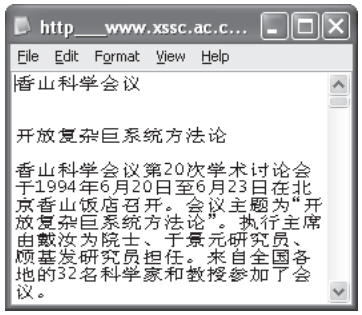

图 5 从 XSSC 页面中抽取出来的会议内容

图 6 从 XSSC 页面中抽取出来的参会者信息

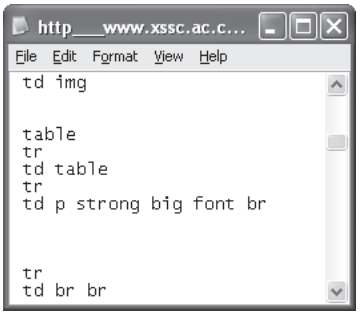

图 7 从 XSSC 页面中抽取出来的 HTML 标签

4.2.3 索引香山科学会议网页

首先,AIS4XSSC 利用 ICTCLAS(http://ictclas.cn/) 中文分词程序对抽取出来的会议内容信息进行了分词处理. 根据 Luhn[21] 的理论,AIS4XSSC 按照图 8 所示的方法从文档中选择符合要求的词语作为 Web 文

本的关键词并对其进行了索引. 虽然目前最为流行的文本关键词选择方法是 TF*IDF 方法[22], 但是实际上 Luhn 方法和 TF*IDF 方法在本质上并无差异, 都是基于统计的方法, 而且, Luhn 方法的计算量却要小很多. 另一方面, 为了利用 Luhn 方法对 XSSC 会议内容进行自动文摘也是本文采用 Luhn 的文本关键词选取方法的动机之一.

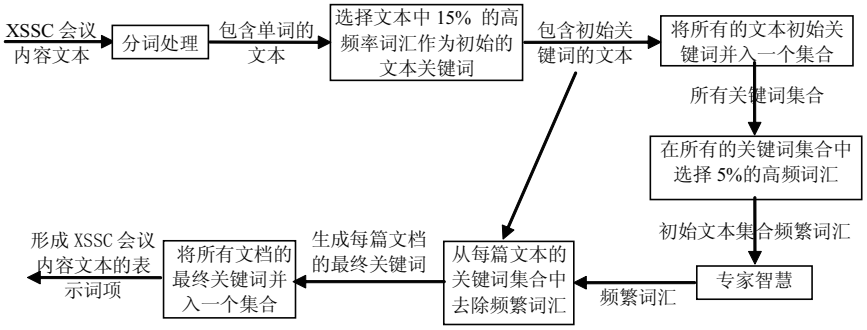

图 8　按照 Luhn 理论设计的香山科学会议 Web 文本关键词选取方法

4.2.4　对香山科学会议网页内容进行文本总结

在对 Web 文本的内容建立起关键词索引之后, 就可以运用 Luhn 的自动文摘算法对文章中的每个句子进行评分了. 然后, 根据一定的压缩比率选出一定数量的高评分句子按照文章中原有的顺序排列作为文章的摘要. 本文中, 对所有的文本内容, 这个压缩比例选取为 0.30. 图 9 是 Luhn 自动文摘句子评分算法.

－ － － [# — — # — # # #] － — — #

－ 表示非关键词;

表示关键词;

[……] 表示设定的句子中词的 Window 的长度,一般根据经验选取，此处取为长度为 8;

Luhn 句子评分公式：Rank ＝ n^2/l

(其中 n 表示关键词的个数，l 表示 Window 的长度)

在上面的例子中，该句子重要性 ＝ 5×5/8 = 3.125

图 9　Luhn 自动文摘句子评分算法

4.2.5　对香山科学会网页进行聚类分析

首先用图 8 中得到的文本关键词集合用布尔表示方法表示每一篇文档集合中的 Web 内容文本; 之后, 利用夹角余弦计算这些文本向量之间的相似度; 然后, 用每一个文本与文本集合中所有文本的相似度来表示这个文本的聚类向量进而利用 SPSS 软件进行层次聚类. 在具体的实现中, 本文选择了 192 篇文档, 因为许多页面的文档内容太短, 不适合向量表示方法. 在合并了这些 192 篇文档的关键词向量过后, 得到了一个 8352 维的向量. 然后用这个 8352 维的向量表示了 192 篇文档, 得到一个 192×8352 维的向量. 然后用夹角余弦计算这 192 个向量之间的相似度, 得到一个 192×192 的对称阵, 最后采用文本之间的相似度进行层次聚类成了 35 个类别. 关于本节香山科学会议 Web 网页聚类的具体技术细节可参考文 [23–24].

4.2.6　对香山科学会议网页进行自动文本分类

首先利用 4.2.5 节中的聚类结果结合香山科学会议的分类标准体系标出各个香山科学会议内容文档所属的类别, 然后采用机器学习算法建立分类器来自动分类文档. 在 AIS4XSSC 系统中, 支持向量机和后向传播神经网络被用来建立统计机器学习的分类器, 实验结果表明, 采用二者的组合学习 (Ensemble learning) 方法将会超过它们任何一种分类算法的效果. 关于此节的具体技术细节可参见文 [25–26].

4.2.7　利用多词 (Multi-word) 技术进行技术术语自动推荐

多词技术是最近计算语言学领域的一个研究热点. 其主要出发点在于捕捉单个词语的上下文信息来增加词语的语义信息. 目前, 它已经在歧义消解, 机器翻译和 OCR (Optimal character recognition) 等方面得到了广泛的应用. 本文研究多词的目的在于从香山科学会议的内容文本中识别出专业的科技术语, 用以帮助用户在使用 AIS4XSSC 系统时推送科技概念, 将机遇关键字的搜索方式提升为基于概念引导的搜索方式. 文

中使用的多词识别程序主要基于 Justeson 和 Katz[27] 提出的技术术语识别方法. 首先识别出文中的重复出现的字符串模式, 然后利用 (1) 式中的正则表达式来截取重复的字符串, 使之符合规范的技术术语词法结构. 在 (1) 式中, A 表示形容词, N 表示名词, P 表示介词.

$$((A|N)^{+}|\ (A|N)^{*}(NP)^{?}\ (A|N)^{*})\ N \quad (1)$$

用于自动识别两个句子中重复的字符串模式的匹配程序如算法 2 所示.

输入:
S_1, 第一个被分词后的句子;
S_2, 第二个被分词后的句子;
输出:
S_1和S_2中的重复字符串模式;
处理过程:
$S_1 = \{w_1, w_2, ..., w_n\}$, $S_2 = \{w_1', w_2', ..., w_m'\}$, k=0
For each word w_i in S_1
  For each word w_j in S_2
    While(w_i equal to w_j)
      k++
    End while
    If k>1
      extract the words from w_i to w_{i+k} to form a repetition
      k = 0
    End if
  End for
End for

算法2 用以识别两个句子中重复字符串模式的匹配程序

4.3 应用效果

图 10 是 AIS4XSSC 系统的用户界面, 以 "复杂" 为关键词进行搜索, 得到了与这个关键词相关的香山科学会议的网页.

图 11 是 AIS4XSSC 系统为用户在进行会议内容检索时提供的两种信息推动功能. 图中的人名推动自动识别了与 "郭雷" 有关的香山科学会议信息; 多词抽取自动推送了香山科学会议中与 "复杂" 有关的技术术语和技术概念. 这部分的功能是最近提出的对象级别的垂直搜索技术 (Object-level vertical search)[28] 和主动推送技术 [29] 在 AIS4XSSC 系统上的具体实现.

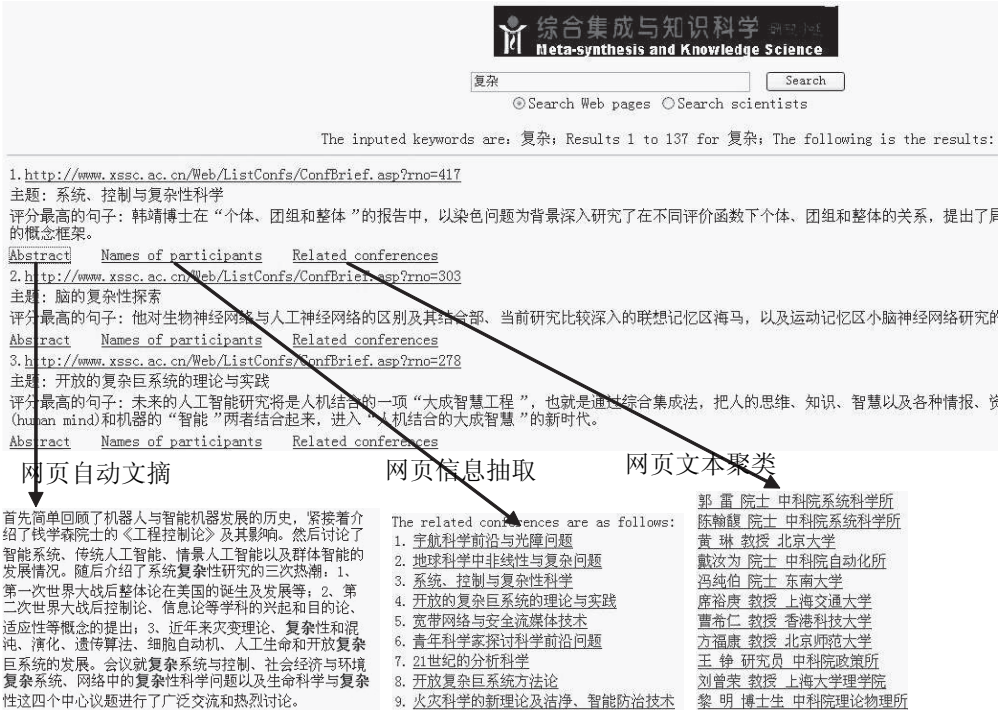

图 10 AIS4XSSC 用于解释 Web 文本挖掘结果的用户界面

5 结论及展望

在综合集成方法论 [30−31] 的指导之下, 本文在当前的网络 "信息爆炸" 形势下以及调研了搜索引擎技术的基础上, 提出了用 Web 文本挖掘的技术去解决有效发现 Web 上有价值信息的知识. 在调研了国内外对于 Web 文本挖掘的学术研究经验之后, 本文提出了 AIS 技术, 并具体分析实现 AIS 技术所需要的各个组件:Web 文本收集,Web 文本索引,Web 文本挖掘处理和用户使用接口界面. 然后, 对于与具体 Web 网站依赖程度不高的部分功能作了探讨. 最后, 基于 AIS 技术, 本文开发了 AIS4XSSC 文本挖掘系统并对诸 Web 文本挖掘技术在香山科学会议 Web 网页挖掘上的具体应用做了详细说明. 具体来说, AIS4XSSC 文本挖掘系统

主要利用了 5 种挖掘技术: 信息抽取, 文本总结, 文本聚类, 文本分类和多词抽取. 限于篇幅的关系, 本文并没有对文本聚类和文本分类做过多的讨论, 而对其他 3 种技术做了详细描述. 最后, 本文展示了 AIS4XSSC 系统针对 XSSC 网站进行了实际应用, 并细致阐述其如何支持用户检索香山科学会议内容.

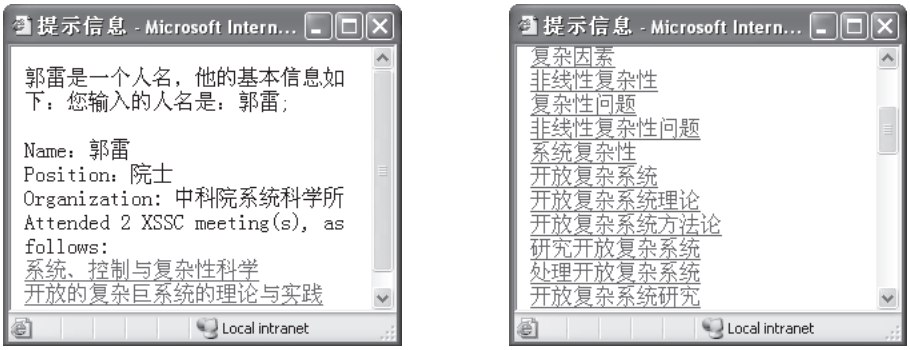

图 11　AIS4XSSC 中的信息推送

下一步的研究将试图提出一个统一 Web 文本挖掘框架, 不仅针对某一个特定的网站, 而且要针对某一类型的网站, 例如新闻, 电子商务网站等站点做进一步的 Web 文本挖掘探索. 例外, 引入更多的计算语言学技术如语义网络 (Semantic Web), 本体技术 (Ontology) 等到 Web 文本挖掘, 提高挖掘的效率和价值也是将来研究的一个关注点.

参考文献

[1] WWW FAQs: How many web pages are there?[EB/OL]. http://www.boutell.com/newfaq/ misc/sizeofweb.html.

[2] White C. Consolidating, accessing and analyzing unstructured data[EB/OL]. http://www.b-eye-network.com /view/2098.

[3] 顾基发, 王浣尘, 唐锡晋, 等. 综合集成方法体系与系统学研究 [M]. 北京: 科学出版社, 2007.
Gu J F, Wang H C, Tang X J, et al. Methodology of Meta-Synthesis and Research on Systems Science[M]. Beijing: Science Press, 2007.

[4] Tang X J. Toward meta-synthetic support to unstructured problem solving[J]. International Journal of Information Technology & Decision Making, 2007, 6(3): 491–508.

[5] Hotho A. A brief survey of text mining[EB/OL]. http://www.kde.cs.uni-kassel.de/hotho/ pub/2005/hotho05Text Mining.pdf.

[6] Hearst M. Untangling text data mining[C]// Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, 1999: 27–56.

[7] Feldman R, Dagan I. Mining text using keyword distribution[J]. Journal of Intelligent Information Systems, 1998, 10: 281–300.

[8] Hotho A, Staab S, Stumme G. Ontologies improve text document clustering[C] // Proceedings of the 3rd IEEE International Conference on Data Mining, Melbourne, Florida, USA, 2003: 541–544.

[9] Yang H C, Lee C H. A text mining approach for automatic construction of hypertexts[J]. Expert Systems with Applications, 2005, 29(4): 723–734.

[10] Lo S H. Web service quality control based on text mining using support vector machine[J]. Expert Systems with Applications, 2008, 34(1): 603–610.

[11] Zhang Z Y, Nasraoui O. Mining search engine query logs for social filtering-based query recommendation[J]. Applied Soft Computing, 2008, 8(4): 1326–1334.

[12] Liu J N K, Dai H H, Zhou L N. Intelligent financial news digest system[C] // Proceedings of KES, 2005: 112–120.

[13] 林鸿飞, 贡大跃, 张跃, 等. 可视化中文文本挖掘模型 [J]. 计算机科学, 2000, 27(4): 37–41.
Lin H F, Gong D Y, Zhang Y, et al. Visual text ming models on Chinese[J]. Computer Science, 2000, 27(4): 37–41.

[14] 刘丽珍, 宋瀚涛, 陆玉昌. Web 使用挖掘的应用研究 [J]. 计算机科学, 2000, 30(9): 46–48.
Liu L Z, Song H T, Lu Y C. Study on application of Web usage mining[J]. Computer Science, 2000, 30(9): 46–48.

[15] 王伟强, 高文, 段立娟. Internet 上的文本数据挖掘 [J]. 计算机科学, 2000, 27(4): 32–36.
Wang W Q, Gao W, Duan L J. Textual data mining on internet[J]. Computer Science, 2000, 27(4): 32–36.

[16] 孙建涛, 沈抖, 陆玉昌, 等. 网页分类技术 [J]. 清华大学学报: 自然科学版, 2004, 44(1): 65–68.
Sun J T, Shen D, Lu Y C, et al. Techniques of Web page categorization[J]. Journal of Tsinghua University: Science and Technology, 2004, 44(1): 65–68.

[17] 王继成, 潘金贵, 张福炎. Web 文本挖掘技术研究 [J]. 计算机研究与发展, 2000, 37(5): 513–520.
Wang J C, Pan J G, Zhang F Y. Study on Web text mining[J]. Computer Research and Development, 2000, 37(5): 513–520.
[18] 鲁明羽, 张红, 付克明, 等. WebME — 一个大型网络挖掘环境系统 [J]. 哈尔滨工业大学学报, 2004, 36(9): 1164–1172.
Lu M Y, Zhang H, Fu K M, et al. WebME — A large scale system on web mining[J]. Journal of Harbin Institute of Technology, 2004, 36(9): 1164–1172.
[19] Wang Z J, Wu J N. A specialized search engine for city classification information retrieval[C]// Proceedings of the 8th International Symposium on Knowledge and System Sciences, Ishikawa, Japan, 2007: 134–139.
[20] Li L Y, Tang S W, Yang D Q, et al. EGA: An algorithm for automatic semi-structured web documents extraction [C] // Database Systems for Advanced Applications, 2004: 787–798.
[21] Luhn H P. The automatic creation of literature abstracts[J]. IBM Journal of Research and Development, 1958, 2(2): 159–165.
[22] Sparck J K. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of Documentation, 1972, 28: 11–21.
[23] Zhang W, Tang X J. A study on web clustering with respect to Xiangshan science conference[C] // Communications and Discoveries from Multidisciplinary Data, Berlin Heidelberg: Springer, 2008: 127–136.
[24] Zhang W, Tang X J, Yoshida T. Web text mining on a scientific forum[J]. International Journal of Knowledge and Systems Sciences, 2006, 3(4): 44–51.
[25] 张文, 唐锡晋. 基于 Web 内容挖掘的信息支持工具 AIS-GAE[J]. 管理评论, 2006, 18(9): 21–26.
Zhang W, Tang X J. Information support tool based on Web content mining: AIS-GAE[J]. Management Review, 2006, 18(9): 21–26.
[26] Zhang W, Tang X J, Yoshida T. Text classification with support vector machine and back propagation neural network[C] // Proceedings of ICCS, 2007: 150–157.
[27] Justeson J S, Katz S M. Technical terminology: Some linguistic properties and an algorithm for identification in text[J]. Natural Language Engineering, 1995, 1(1): 9–27.
[28] Nie Z, Wen J, Ma W. Object-lelvel vertical search[C]// Proceedings of 3rd Biennial Conference on Innovative Data Systems Research, 2007: 235–246.
[29] Hermans B. Information push and information pull[EB/OL]. http://www.hermans.org/agents2/ch3_1_2.htm.
[30] 唐锡晋. 综合集成研讨厅的几个示例 [J]. 系统科学与数学, 2009, 29(11): 1507–1516.
Tang X J. Some examples of ther HWMSE[J]. Systems Science and Mathematics, 2009, 29(11): 1507–1516.
[31] Tang X J, Liu Y J. Zhang W. Augmented analytical exploitation of a scientific forum[C] // Communications and Discoveries from Multidisciplinary Data. Berlin Heidelberg: Springer, 2008: 65–80.