文章编号:1001 - 9081(2009)06 - 1601 - 04

基于语义 Web 服务的分布式服装搜索引擎系统设计

张革伕,徐 琪
(东华大学 旭日工商管理学院,上海 200051)
(gzjkfu@mail.dhu.edu.cn)

摘 要:从电子商务环境下服装供应链管理的需求出发,分析了目前服装搜索引擎存在的问题,提出了基于语义 Web 服务的分布式服装商品搜索引擎系统模型,并讨论了它的体系结构。介绍了基于 Ontology Web Language (OWL)的服装本体设计模型及其语义描述方法。分析了服装搜索引擎的基本功能及分布式环境下的 Web Services (WS)合成。理论分析和实例原型说明了基于服装语义树的搜索引擎多关键词搜索效率明显高于全文搜索引擎。

关键词:语义 Web; Web 服务; 分布式搜索引擎; 服装供应链
中图分类号: TP39 **文献标志码**:A

Design of distributed search engine system for apparel based on semantic Web services

ZHANG Ge-fu, XU Qi
(*College of Glorious Sun Management, Donghua University, Shanghai* 200051, *China*)

Abstract: According to the needs of supply-chain management in the environment of E-business, the authors analyzed the problems existing in the present search engine for apparel-goods and proposed a new distributed search engine for apparel goods based on semantic Web Services (WS), and then discussed its architecture. After introducing the apparel Ontology design model and the representation method based on Ontology Web Language (OWL), the basic functions of apparel search engine and WS synthesized in distributed environment were presented. Theoretical analysis and the designed prototype show that this search engine, based on apparel semantic tree, is much more efficient than full-text search engine when searching with multi-keywords.

Key words: semantic Web; Web Services (WS); distributed search engine; apparel supply-chain

0 引言

电子商务的快速发展为服装供应链的上下游成员带来了巨大商机的同时,也带来了管理与技术上挑战。服装供应链如何快速响应复杂多变的市场需求已经成为电子商务迫切需要解决的一个重要课题。一方面,客户在纷繁复杂的市场中,面对海量的产品信息数据,无法快速准确地选择到自己满意的服装;另一方面,面对快速多变的服装市场形势,供应链成员尤其是销售商缺乏快速推销服装给客户的有效技术手段,以至失去商机,增加库存及销售损失。

电子商务环境下,搜索引擎既是让消费者快速找到所需求商品的最有效方法之一,也是让商家快速将商品推向市场的最佳技术之一。但是目前的综合型搜索工具搜索到的结果往往是非需求信息多得把主题信息都湮没了[1],更严重的是这些工具不能准确搜索到客户所需的服装。国内外有关专业服装搜索引擎研究与开发的文献还不多,与服装搜索相关的引擎有两种:一种是 B2C 形式的服装零售商铺搜索(例如:http://www.souyifu.com/product_list.aspx);另一种是 B2B 形式的服装生产商或贸易商目录搜索(例如:http://www.apparelsearch.com/search_engines.htm)。所有工具的实现都不以快速定位服装为目标,不涉及具体的商品形态特征,使得技术上无法保证供应链快速响应市场。这是因为现有搜索引擎既不能理解用户输入信息的意义,也不能分析搜索结果与需求间的吻合程度。

语义搜索是解决当今搜索问题的一个新思路[2],可用机器理解输入输出的技术来解决传统服装搜索引擎缺陷。这种技术集成了新的语义 Web (Semantic Web)、Web 服务(Web Services,WS)等多种技术,为开发有效的服装搜索引擎提供了技术保障。

1 语义 Web 服务

语义 Web 服务严格来说,包括两个内容:语义 Web 和 Web 服务,语义 Web 是更高效地实现 Web 服务的基础设施。

语义 Web 是互联网创始人 Tim Berners-Lee 在 1998 年提出的,认为要解决目前计算机对信息的无知问题,需要在现有 Web 的基础上将资源人为地赋予各种明确的语义信息[3]。为使分布式和异构网络环境下的机器能够相互理解,语义 Web 的页面内容来自使用 XML 标记的元数据,也就是内容信息的存储及其表示结构使用的是 XML 标签形式;并用本体(Ontology)语言定义 XML 标签的语义。在一个领域内,本体包括 2 个内容:概念与规则(关系),本体就是指对概念化对象的明确表示和描述,概念和规则一起构成了领域知识[4]。

Web 服务是一种部署在 Web 上的对象,是一个暴露于外界的、能够通过因特网进行调用的 API,或者说是应用程序,

收稿日期:2008 - 11 - 28;**修回日期**:2009 - 02 - 11。 **基金项目**:国家自然科学基金资助项目(70772073);上海市自然科学基金资助项目(07ZR14003);上海市社科规划基金资助项目(2007BZH001)。

作者简介:张革伕(1969 -),男,湖南益阳人,博士研究生,主要研究方向:供应链管理、计算机网络; 徐琪(1963 -),女,浙江台州人,教授,博士生导师,主要研究方向:系统集成、供应链管理。

极大地实现了软件代码及其功能的复用。良好的封装性、松散偶合、使用标准协议规范和高度可集成能力成为其备受推崇的理由，其实现依赖于XML、SOAP、WSDL和UDDI四个技术，主要包括：数据存储、交换、功能描述与发布、发现和调用等过程[5]。

Web服务并没有为机器理解提供任何机制，而语义Web的语义标注无疑有助于丰富和扩展现行Web服务语法层级上的功能。研究表明将Web服务技术与语义Web技术糅合，可以大幅度地提高用户对Web服务的发现、调用、组合的自动化和智能化程度。现在，语义Web服务把Web服务和业务理解有机地结合起来，为解决复杂业务过程的协调与敏捷性提供了新思路[6]。

2 分布式服装搜索引擎系统架构

分布式服装搜索引擎的目标是让用户快速、准确地找到自己想要的服装商品，甚至实现个性化服装的组合搜索服务，最后迅速完成交易。其分布特征首先表现在供应链成员的信息平台及其提供的用户操作接口的分布性；其次，数据来源具有物理位置的分散性和自治性，各个平台拥有者无须为专有数据的可能暴露担心。图1为分布式搜索引擎的体系结构，供应链上所有的成员在因特网上都处在一个对等的地位，按照一定的合作策略，任意用户接口程序通过调用各合作伙伴暴露在外的搜索Web服务(SWS)来读取商品信息库(XML)。通过这种机制，分散且自治的供应链网络各成员构成了一个市场快速响应服务联盟。

在分布式网络环境下，每个企业可能使用了不同的数据库表来存储信息。这种结构存在两种局限性：一是没有存储服装特征及其之间的语义，造成不同厂商之间的数据交换可能存在障碍；二是由于服装属性特征众多，这种结构对服装这种特殊类型的商品的存储效率低，并且多关系表的关联搜索效率也很低。为解决上述问题，本文使用OWL(Ontology Web Language)来定义和描述服装本体，按照语义上的层次与约束关系来存储数据，得到服装本体的XML定义文档，进而得到高效率的服装语义搜索树。

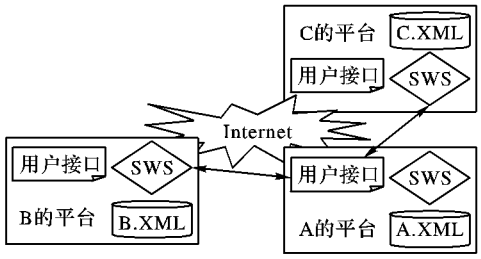

图1 分布式搜索引擎体系结构

考虑一个实际的供应链网络，设其上有n个生产商、m个销售商拥有自己的基于Web的网络信息管理与交易平台，1个销售商可从多个生产商订购商品。每个销售商从生产商采购服装后，按照预先的服装本体特征生成一个XML格式的商品清单，清单包含服装商品的种种特征和双方交易地址URL等信息，服装搜索引擎服务程序按照输入信息的语义特征来读取信息。基于语义Web服务的服装搜索引擎系统框架如图2所示。在图2中，UDDI是Web服务注册、查询服务中心，可设立在第三方服装评价与流行预测中心，为所有服装供应链网络上的成员提供搜索服务SWS的注册、查询、绑定以及证书服务，而且只有在UDDI中注册的WS才能进行分布式环境下的合成；预处理模块根据语义树来管理服装的商品信息，并生成XML数据库，其预处理程序以WS的形式提供给所有合作成员调用；服装本体库为语义树提供概念对象集及其规则集。

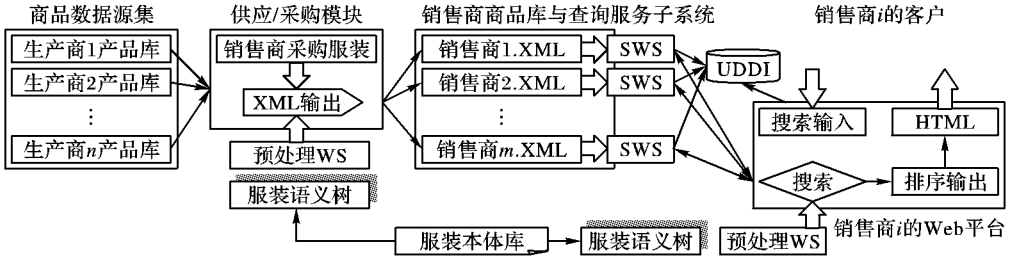

图2 基于语义Web服务的服装搜索引擎系统框架

3 服装本体及其搜索Web服务

OWL称为Web本体语言，是W3C确认并推荐的语义Web描述规范的一个部分，包含三个子语言：OWL Lite、OWL DL、OWL Full。OWL被设计成用于处理Web站点信息的内容，通过提供更多的语义解释，使得语法结构层次上的XML内容有了机器可理解性。下面将介绍服装本体的形式化描述，以及基于语义的服装SWS功能架构。

3.1 基于OWL的服装本体模型

在服装领域内，利用OWL Lite来内定义概念、描述概念之间的约束关系，首先需要抽象出类及其属性，可构造出图3所示的服装本体的类及其属性模型，图3中WearableThings为抽象类，女性半身裙FemaleSkirt、男性衬衫ManShirt等基本款都为其子类，并且互为兄弟关系；款式名TypeName为ManShirt等的属性(property)，属性还可以有自己的子属性(subproperty)。下面给出服装本体基于OWL Lite的概念及其关系描述片段。

```
<owl: Class rdf: ID = "WearableThings"/ >
<owl: Class rdf: ID = "ManShirt" >
    <rdfs: subClassOf rdf: resource = "#WearableThings"/ >
</owl: Class >
<owl: Class rdf: ID = "Garment-Mshirt" >
    <rdfs: subClassOf rdf: resource = "# ManShirt "/ >
</owl: Class >
<owl: Class rdf: ID = "BrandGarment-Mshirt" >
    <rdfs: subClassOf rdf: resource = "#Garment-Mshirt"/ >
        <rdfs: subClassOf >
        <owl: Restriction >
            <owl: onProperty rdf: resource = "#madeWithBrand"/ >
        </owl: Restriction >
        </rdfs: subClassOf >
    </owl: Class >
<owl: ObjectProperty rdf: ID = "madeWithBrand" >
    <rdfs: domain rdf: resource = "#GarmentBrand-Mshirt"/ >
    <rdfs: range rdf: resource = "#Brand"/ >
</owl: ObjectProperty >
<owl: DatatypeProperty rdf: ID = "HasPrice " >
```

```
< rdfs: domain rdf: resource = "#SaledGarment-Mshirt" / >
< rdfs: range rdf: resource = "&xsd;  decimal"/ >
</owl: DatatypeProperty >
```

通过定义标签,使得存储商品信息的XML文档意义明确,并通过关系约束、定义域与值域的规范,概念及其属性特征在语义描述上得到了统一,例如处理“异名同义、同名异义”,从而实现异构环境下的数据交换,也能使各成员平台对用户输入有相同的响应方式。

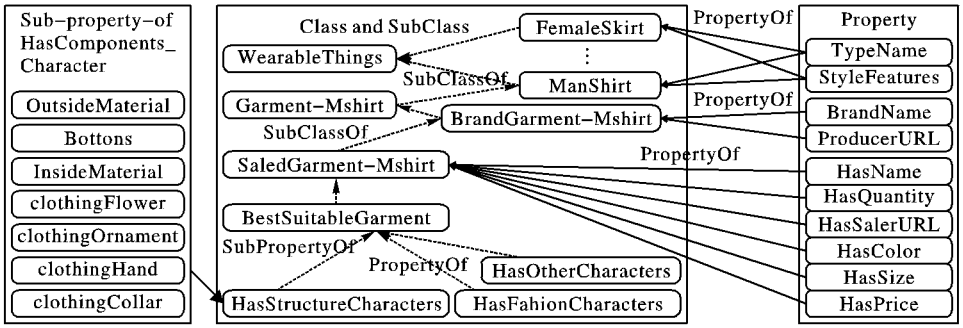

图 3　服装本体的类及其属性模型

3.2　服装语义树

根据前面的语义定义,销售商的商品数据存储结构为如图 4 所示的服装语义树,文件格式基于 XML。由于文件所使用的 XML 标签都有明确的定义,其取值有严格的约束,因此数据的意义明确、冗余小。显然,服装语义树具有严格的层次性,被分成了范围大小不同,但意义不重叠的域。由于在用户输入服装搜索信息时,本身就包含明确的语义,系统容易根据服装本体特征,解析出语义明确的关键词,通过语义树的引导,逐步缩小搜索域,实现高效率搜索。

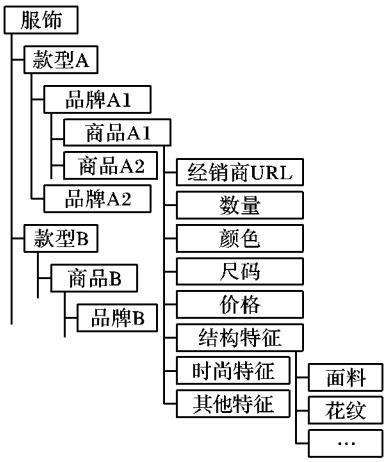

图 4　基于 XML 的服装商品语义树

3.3　搜索 Web 服务及其合成模型

对于搜索服装,尽管用户的搜索请求输入是随意的,但语义清楚。首先,输入内容必定和服装款型有关;其次,输入内容本身具有语义约束特征,例如:“白色 丝绸 连衣裙”,通过加入语义明确的限定词以锁定服装商品。根据搜索目的,用户的搜索请求语义可以简单归纳为:

1)根据款型,搜索各种品牌服装以进行对比;

2)根据款型和品牌,搜索足够数量的服装品种以帮助选择;

3)根据服装品牌、款型以及市场信息和时尚与服装结构特征等信息,搜索最合适自身需求特征的服装。

由于合作的需要,将这些功能设计成 WS,以供合作各方 Web 平台调用。为完成复杂的功能,需要将这些独立的 WS 进行合成。基于语义的 WS 合成技术,在文献[5]中有详细地论述,图 5 所示的服装本体搜索 WS 模型描述了这种合成逻辑。下面定义搜索 Web 服务。

用户通过搜索引擎接口输入信息时,预处理程序 WS(PWS)通过分析其语义,根据服装本体库及其语义树来产生一个 n 元组,这个 n 元组是 SWS 的输入参数。

定义 1　设输入为 PI_i,则 PWS 的输出为一个 n 元组:$\langle I_1, \cdots, I_n \rangle$,其中:$n \geqslant 1$,满足 $\langle I_1, \cdots, I_n \rangle = PWS(PI_i)$。

定义 2　如果 SWS 的输入为一 n 元组:$\langle I_1, \cdots, I_n \rangle$,SWS 的输出为一 m 元组 $\langle O_1, \cdots, O_m \rangle$,其中:$n \geqslant 1$、$m \geqslant 0$,WS 则满足:$WS_i(I, O) = SWS_i(\langle I_1, \cdots, I_n \rangle; \langle O_1, \cdots, O_m \rangle)$。

下面 1)~4)定义了四个 SWS,分别为:SearchByType、SearchBy Brand &Type、SearchBySaledList、Search _ BestBy Characters。其中:〈[HasSize],[Ha_ sColor], [HasPrice] 〉表示此元组中的三个属性值至少有一个出现,例如用户可能因只关心颜色而只键入颜色描述词,也可能只在乎价格,或者三个元素都必须同时考虑,[HasPrice]表示可以为空值;同样〈[FashionCharac ters],[StructureChar acters],[OtherCharacters]〉表示此元组中的三个元素至少应出现一个,每个元素本身也可为一个元组。

1) SearchByType(〈TypeName〉; SaledGarment)

2) SearchByBrand&Type(〈TypeName, [BrandName]〉; SaledGarment)

3) SearchBySaledList(〈TypeName, [BrandName], 〈[HasSize],[HasColor],[HasPrice]〉〉; Garment)

4) SearchBestByCharacters (〈TypeName, [BrandName], 〈[HasSize], [HasColor], [HasPrice]〉, 〈[FashionCharacters], [StructureCharacters], [OtherCharacters]〉〉; Garment)

但是,分布式环境下,WS 的合成还存在业务流程上的问题,例如:决定串行、并行执行搜索顺序,如何处理一方的搜索失败。BPEL4WS 是近年来提出的面向工作流的合成标准,BPEL4WS 定义了销售商平台内搜索流程之间以及与其合作伙伴之间使用 WS 来进行交互的情况,一个可执行流程本身就是一个 WS,接口用 WSDL 文件来描述。文献[5]介绍了一种基于最小执行代价的自动合成算法,本研究对该算法条件进行对应修改:假定销售商 A 对每个合作者进行优先级选择,优先等级值域为[0, ∞)内的正整数,数值越大,优先级越低。最小值 0 为最高优先级,指销售商 A 自身。

4　分布式服装搜索引擎实现实例

东华大学供应链快速响应策略研究组根据前面的模型,设计出分布式环境下、基于语义 Web 服务的服装搜索引擎系统原型。该系统中,为了避免用户搜索时需要输入过多的信息,采用 Ajax 异步机制,让用户在输入一个关键词后自动给

出最大可能出现的下一个搜索语句集,用户只需要在这些语句后键入空格就可以继续显示下一个可能的搜索语句集,以帮助用户快速匹配自己的搜索目标并减少用户输入工作量。这种模式对供应链成员统计服装流行元素的构成非常快捷。图6为搜索结果的界面图,图的上部分为基于Ajax模式的用户输入框,图的下部分为选择"夏奈尔 连衣裙 粉色"后所精确搜索到的商品。

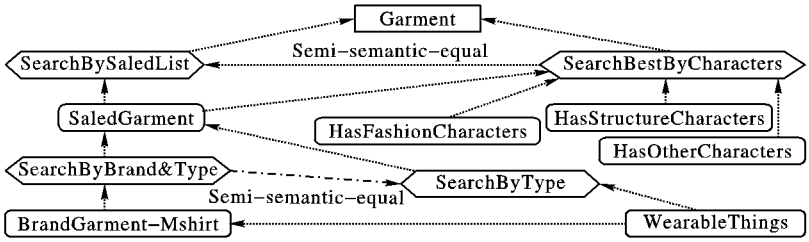

图5 服装本体的SWS模型

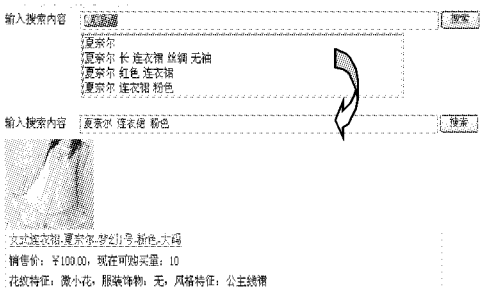

图6 服装搜索引擎搜索结果的界面

与目前实际使用的搜索引擎相比,本文所介绍的引擎的搜索结果更加精确和高效,呈现出良好的性能特征:1)基于语义树的搜索模式通过逐步缩小搜索域,从逻辑层次上实施多关键词搜索,其算法效率要高于全文搜索模式;2)支持合作伙伴之间因业务需要而产生的搜索,可以灵活地根据协作策略规定搜索结果的排列顺序;3)原则上,本服装搜索引擎支持语义明确的任意长语句输入搜索形式。

5 结语

服装搜索引擎与目前通用的搜索引擎相比,由于使用了明确的语义定义XML标签,能让搜索程序充分理解概念的意义以及概念之间的关系,在分布式环境下,表现出良好的数据理解适应性。对于消费者而言,搜索引擎能根据消费者的个性化穿着需求,在整个行业内进行搜索,以最大限度地满足需求;对于供应链成员,需要在语义上理解消费者的需求特点和变化,协同操作,快速响应市场需求。这种分布式搜索引擎系统为电子商务背景下的供应链协同运作提供新的工具。由于目前在不同平台的数据表达统一预处理方面采用的仍然是穷举推理方法,灵活性受到限制,进一步的研究方向是智能处理方法在服装搜索引擎中的应用。

参考文献:

[1] SHU B, KAK S. A neural network-based intelligent metasearch engine [J]. Information Sciences, 1999, 120(1/4): 1 - 11.

[2] CNET 科技资讯网. Google 面临新技术挑战,语义搜索潜力大[EB/OL]. [2008 - 06 - 25]. http://www.cnetnews.com.cn/2008/0618/933251.shtml.

[3] LEE T B, HENDLER J, LASSILA O. The semantic Web [J]. Scientific American, 2001(5): 34 - 43.

[4] STUDER R, BENJANMINS V R, FENSEL D. Knowledge engineering, principles and methods [J]. Data and Knowledge Engineering. 1998, 25(1/2): 161 - 1971.

[5] 顾宁,刘家茂,柴晓路. Web Services 原理与研发实践[M]. 北京:机械工业出版社,2007: 1 - 11.

[6] 宋庭新,黄必清,熊健民,等. 语义 Web 服务在业务协同与供应链集成中的应用[J]. 中国机械工程,2008,19(4): 410 - 413.

(上接第1600页)

案例库与维护前的案例库在分类准确率上基本相当。

参考文献:

[1] SMYTH B. KEANE M T. Remembering to forget: A competence-preserving case deletion policy for case-based reasoning systems [EB/OL]. [2008 - 10 - 10]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.2767.

[2] FRANCIS A G, RAM A. The utility problem in case-based reasoning [EB/OL]. [2008 - 10 - 10]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.639.

[3] 耿焕同,肖明军,邹翔,等. 聚类算法在范例库维护中的应用研究[J]. 计算机工程,2005,31(12): 166 - 168.

[4] 鲍玉斌,王琢,孙焕良,等. 一种基于分形维的快速属性选择算法[J]. 东北大学学报:自然科学版,2003,24(6): 527 - 530.

[5] LEE H D, MONARD M C, WU FENG-CHUNG. A fractal dimension based filter algorithm to select features for supervised learning [EB/OL]. [2008 - 10 - 10]. http://www.icmc.usp.br/~iarn2006/sbia/apresentacoes/IBERAMIASBIA_TS1/IBERAMIASBIA_TS1_A2.pdf.

[6] BARBARA D, CHEN P. Using the fractal dimension to cluster datasets [C]// ACM-SIGKDD: Proceedings of the 2000 International Conference in Knowledge Discovery and Data Mining. Menlo Park, California: AAAI Press, 2000: 260 - 264.

[7] HIPPENSTIEL R, EL-KISHKY H, RADEV P. On time-series analysis and signal classification - part I: Fractal dimensions [C]// Conference Record of the 38th Asilomar Conference on Signals, Systems and Computers. Washington, DC: IEEE Press, 2004, 2: 2121 - 2125.

[8] BARBARA D, NAZERI Z. Fractal mining of association rules over interval data [R]. Fairfax: George Mason University, 2000.

[9] 闫光辉,李战怀,党建武. 基于多重分形的聚类层次优化算法[J]. 软件学报,2008,19(6): 1283 - 1300.

[10] BARBARA D. Chaotic mining: Knowledge discovery using the fractal dimension [EB/OL]. [2008 - 10 - 10]. http://www.isse.gmu.edu/techrep/1999/99_03_barbara.ps.

[11] BORGNAT P, FLANDRIN P, AMBLARD P O. Stochastic discrete scale invariance [J]. IEEE Signal Processing Letters, 2002, 9(6): 181 - 184.

[12] MAZEL D S, HAYES M H. Using iterated function systems to model discrete sequences [J]. IEEE Transactions on Signal Processing, 1992, 40(7): 1724 - 1734.

[13] 张济忠. 分形[M]. 北京:清华大学出版社,2001.

[14] 姜灵敏,周锋. 上证指数盒维数的计量与特性研究[J]. 系统工程学报,2006,21(4): 434 - 437.

[15] 倪志伟,李锋刚,毛雪岷. 智能管理技术与方法[M]. 北京:科学出版社,2007.

[16] 倪丽萍,倪志伟,吴昊,等. 基于分形维数的数据挖掘技术研究综述[J]. 计算机科学,2008,35(1): 187 - 189.

[17] CHANG C C, LIN C J. LIBSVM — A library for support vector machines [EB/OL]. [2008 - 10 - 10]. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[18] Christos Faloutsos — Released software [EB/OL]. [2008 - 10 - 10]. http://www.cs.cmu.edu/~christos/software.html.