Web Community Discovery

信息学院第二届
学科建设与科研工作会议

李翠

计算机科学与技术系

Outline

- Field of study
- Summarize past work
- Enounce present work
- Design future work

Field of study

Field of study: Web Mining and IR

Subject : research and application of web community discovery based on hyperlink analysis

Question?

- Web数据的复杂性：分布式，不稳定，海量，半结构或无结构及冗余，信息质量无保证，异构……
- Web数据的迅猛增长
- 现有Web搜索系统在信息覆盖率、及时性、个性化、可扩展等方面存在的问题
- 促进了Web挖掘、智能信息检索等新兴学科及技术的发展。

Study and purpose: Web community discovery

- Web community: 在Web中有明显关联性、相互之间存在大量链接的Web页面集。
- Web community discovery: 从Web拓扑结构的角度进行网页的聚类。
- 研究基础：链接分析。
- 逻辑上将Web看作为图，Web图G=(V,E)，V节点集，E链接。
- 从节点入度、出度、传递性、聚团等特性出发，分析Web社区结构，研究其发现算法，以实现对Web信息集合的有效划分。使网页返回结果更加相关，优化排序质量。

陕西省自然科学基础研究计划项目 (2007F52)

- Project name: 面向网络资源搜索的Web社区发现算法研究
- Purpose:

①深入研究当前Web信息搜索技术及其系统中的数据资源发现与利用的基本问题;

②设计有效的面向网络资源搜索的Web社区发现算法，结合网络资源搜索等技术，开发Web社区发现引擎;

③深入研究社区中的结构关系，结合现有关键技术和推理规则，对Web信息集合进行有效的主题分类与直观易懂的描述，达到网络环境下信息资源的广泛共享、高质量获取、有效分类等目的。

Content

①Web社区发现算法研究

②主题相关性判定模型研究

③Web社区发现对信息获取策略的改进研究

④Web社区发现对信息检索结果自动分类研究

续

- **Adv.:**

  ①**Web**社区拓扑结构与内容的映射关系.
  ②通过**Web**社区发现算法实现信息的分类.

续

- **Experiment & feasibility**

①结合现有的编程技术及搜索引擎软件工具包所提供的可扩展接口，鉴于Web社区发现算法在搜索引擎中应用能够带来的有效性，建立其实例系统—Web社区发现引擎 。

②如何构造出社区的层次结构，如何抽取出社区的特征表示以及如何准确测量并挖掘出社区之间的关联关系，如何设计适合于Web社区环境下的实用有效的知识发现算法，特别是增量学习与动态学习算法，如何通过知识转换与信息组合等技术实现创新等都是研究关键。

③基于理论研究成果，采用先进的软件设计方法设计核心算法与计算机程序，反复实验，不断改进，从而可提供基于社区的创新型Web信息检索原型工具—Web社区发现引擎。

Past work

- **<1>:The Improved Model of Web Information Search Based on Semantic**

（Journal of Northwest University. 2007.1）

续

- **Aim:** A improved model of Web Information Search based on Semantic is presented in view of current Web information search model's disadvantages.
- **Methods:** Fuzzy association technology and Ontology and Web mining technology is employed.
- **Results:** Model of semantic information search based on five-element-array is built ,and algorithm is given.

$$M_S = [D, O, U, F, R(D \rightarrow U, SemSim(U, D))]$$

①D为文档模型，D=<Dc,Df>;②O为领域模型O=<C,R>; ③U为用户模型U=<Uc,Uf>; ④ F是一个框架，用于用户、领域、文档及三者间关系的动态建模; ⑤ R(D→U,SemSim(U,D))是获得用户信息需求相关文档的算法.

续

<2>:Classification Algorithm and Application of Neural Network Based on CRBF

(Computer Engineering,2007.9)

续

神经网络分类技术的现状与应用：

近年来，神经网络分类分析技术有较快发展。它在模式识别、语音识别、医疗应用、故障检测、问题诊断、机器人控制及计算机视觉等领域得到了很好的应用。使得企业从生产管理的海量数据中挖掘出有用的数据,并为企业提出有用的“建议”,使企业的效益有显著的增加。

神经网络分类算法的实质：

用户根据实际应用选择合适的数学模型，由建立的数学模型对实际输出结果进行分类。

研究出发点

在**RBF(Radial Basis Function)**基础上，试图从提高梯度下降学习效率和修正模型参数，确保分类结果中属性合理性出发，提出了基于**CRBF(Cosine Radial Basis Function)**神经网络数学模型、几何模型、分类算法及应用模型。

过程:简化**RBF**→ $\phi_j(x) = g_j(x^2)$ → 发生器函数→选择参数 → 得**CRBF**模型

$$g_j(x) = \frac{a_j}{(x + a_j^{\,2})^{1/2}} \qquad \text{即} \quad \phi_i(x) = \frac{a_j}{(x^2 + a_j^{\,2})^{1/2}}$$

$$h_{j,k} = g_j(\| x_k - v_j \|^2)$$

<3>:陕西省教育厅专项科研计划项目 (07JK339)

- Project name:面向帐号异动安全监控的混合神经网络核心算法研究
- Purpose:

  ①以粗糙集和概念格理论为基础，建立合理的混合神经网络应用模型，包括其数学模型、几何模型、挖掘和分类算法及训练模型；

  ②以神经网络思想为指导，综合运用**Web**挖掘技术、模糊计算技术及人工智能等技术，设计面向异动帐号监测的实用有效的知识挖掘算法；

  ③综合运用信息组合、知识转换及逻辑推理等方式揭示公有账号到个人账号频繁的、大量的、隐蔽的外汇资金流动的关联。
- Content & technology:

  ⑴用于监测帐号异动的混合神经网络数学及几何模型

  ⑵用于监测帐号异动的信息挖掘模型

  ⑶用于监测帐号异动的混合神经网络训练模型

- Adv.:

(1)提出基于粗糙集、概念格等理论实施对帐号异动的信息表示与知识组织，从而为设计有效的异动帐号挖掘和训练算法打下良好的基础。

(2)提出一种综合运用数据挖掘、机器学习及多种数学计算理论的混合神经网络模型来描述帐号异动的过程，采用间接频度及相似法追踪多个异动帐号共同的资金流向和资金来源。

(3)运用量子理论中叠加思想对我们提出的模型进行合成训练，设计实用有效的训练算法。

(4)开展软聚类等技术进行自动监测帐号异动的创新算法研究。

<4>:Dynamic load Balance Arithmetic supporting Distributed Cooperation Intrusion Detection

（computer Engineering.2008.1）

- Based on the analysis and study of distributed cooperation intrusion detection system and its flow dynamic distribution question, a algorithm of dynamic load balancing is proposed. It hashes the destination ip address in packet to map the corresponding packet to scope of the network data collect agent or detection agent's number, and adjusts the scope according to the performance and load of the network data collect agent or detection agent.
- DCIDM（Distribute Cooperate with Intrusion Detect Modal）and algorithm of dynamic load balancing

算法的基本思想

给每个网络数据采集代理和检测代理设置一个接收数据包的子区间，通过区间的调整来动态的分配网络流量，对某个网络数据采集代理来说，如果把数据包的特征域经过哈希运算后的值在其区间内，则对该数据包作进一步的处理，然后传给检测代理，否则丢弃；某个检测代理接收到数据后也先作同样的处理。这样就充分的利用了系统资源，提高了整体效率。

<5>:Research and Application of Reticulate Clustering Algorithm Based on Auto-structure

(Computer Engineering & Design . 2008.7)

- 聚类可使得物理或抽象的对象划分成具有一定意义的子类，做到不同子类中的对象尽可能相异，而同一子类中的对象尽可能相似。
- 它在信息系统分析、人工智能、决策支持系统、知识发现、模式识别与分类、故障检测、信息检索、基因分析、客户关系管理、网络信息安全等领域得到了广泛的研究与应用。
- 聚类分析的主要方法包括：Partitioning Methods，Hierarchical Methods,density-based methods,grid-based methods,Model-Based Methods……

续

- 实质就是对于**n**个数据项，找到一个最好的划分，以使所有数据项划归到**k**个不重合的组中。即需要定义聚类划分的目标函数。
- 本文主要以基于距离的目标函数为基础并对其进行扩展，将以最小化结点连接代价作为目标函数，构建簇心全连通FCCH（fully-connected cluster heart）和簇心连通CCH（connected cluster heart）两种模型进行数据集的聚类分析。

续

- 自构形网状聚类算法主要思想:

基于自构形网状聚类的主要思想是：对数据进行分级聚类，即通过构建多类向量机**SVM**实现对普通结点层和聚类中心（簇心）层的分级数据处理，部分结点作为簇心，各簇心进行连接形成簇网中心，每个普通结点再与簇心连接，这样形成一个聚类网。自构形就是对簇心拓扑结构和学习算法实现选择和适应的过程，包括解决问题方法的自构形（样本数据几何形状）、簇心层的自构造和算法的自主性三方面。

Present work

- English papers
- Basal knowledge
- Summarization

Future work

- Papers
- Project
- Study

Thanks !