C-DBLP：中文文献信息集成系统

陈威　　王仲远

C-DBLP（http://www.cdblp.cn）是一个以作者为中心的面向计算机领域的中文文献集成系统，由本实验室 Web 组开发完成，于 2008 年 10 月中旬正式对外发布。该项目是实验室承担的国家 863 课题“海量数据空间模型、索引与查询技术研究”（项目编号 2007AA01Z155）下的一个子项目，是我们在个人数据空间中 Web 上个人信息管理机制的一次尝试。在 C-DBLP 系统开发过程中，我们形成了一个可扩展的软件平台，可用来管理具有丰富数据的学术社区（Community）。通过对计算机领域研究人员这个社区的文献信息的分析和处理，我们在学术社区信息管理（Academic Community Information Management）方法上也取得了一定的研究进展。

Ⅰ 系统简介

计算机科学文献库 DBLP Computer Science Bibliography 在学术界有很好的声誉，给人们带来了极大的便利，其权威性也得到了研究界的高度认可。然而 DBLP 不提供对中文文献的收录和检索功能，国内的权威期刊及重要会议的论文缺乏一个类似的集成检索系统。

WAMDM 实验室自 2000 年开始研究 Web 数据集成的相关技术，先后在 Web 数据抽取、数据库选择、查询转换等方面积累了丰富的研究工作和技术成果，并一直在尝试利用研究成果去解决人们在 Web 使用中面临的问题。针对中文文献缺乏权威的收录和检索系统的现状，我们尝试着在计算机领域中建立一个类似于 DBLP 的文献集成系统。2008 年暑假中，我们利用 WAMDM 实验室积累下来的 Web 集成技术，短短几个星期就成功搭建了系统原型，高质量地集成了我国计算机科学领域 11 本权威期刊自创刊以来及中国数据学术会议（NDBC）2000 年至 2008 年来共 5 万余篇文献，提供基于作者的文献检索服务，并在此基础上开展了社区信息管理的研究。这就是面向计算机领域的中文文献集成系统 C-DBLP。

通过两个月的试运行，实验室在 2008 年 10 月中旬正式对外发布了 C-DBLP 系统，并根据用户反馈不断改进系统功能。我们的工作得到了研究界高度认可，中国计算机学会网站和《中国计算机学会通讯》都登载了该系统的发布报告（如图 1 所示）。

图 1 C-DBLP 的系统发布新闻

系统对外发布后，我们还得到了来自广大用户的肯定。截至 12 月初，搜索引擎中 Google 已收录本站近 7 万页面，百度也已收录本系统近万条记录。C-DBLP 网站的访问量反映了我们的工作得到了许多用户的关注，从系统发布以来的用户访问情况（如图 2 所示）可以看出，越来越多的用户开

始关注 C-DBLP，系统网站的访问量总体上不断上升，到 12 月上旬系统平均每日向近两千位用户提供文献检索服务。我们相信，随着系统功能的演进，C-DBLP 将会得到更多的关注和认同。

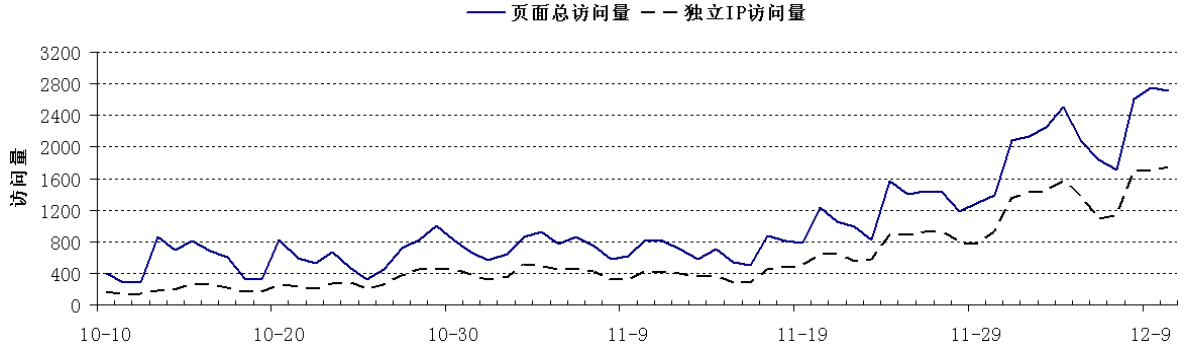

图 2 C-DBLP 系统发布以来用户访问量情况

Ⅱ C-DBLP 与学术社区信息管理

在传统的个人数据管理场景中，本地计算机是存储和处理个人信息的主要载体，如何在桌面电脑中有效管理个人数据是研究界最关注的问题。然而随着 Web 提供越来越强大的功能和更好的用户体验，blog 等面向个人的应用不断涌现，更多的用户选择 Web 作为发布和处理个人信息的平台，这使包含海量的个人信息的 Web 在人们的工作和生活中扮演着越来越重要的角色。区别于本地计算机中个人信息的有限性和规则性，Web 上的个人数据具有高度的异质性和更大的数据量，Web 场景下个人数据的发现和管理给研究人员带来了不小的挑战。针对这一现状，WAMDM 实验室在 Web 上的个人数据管理方面进行了一些探索性的研究。

根据用户行为和偏好分析，我们发现 Web 上存在着许多的社区（Community），这些社区由成员（People）和存储该社区相关信息的多个数据源（Data Sources）组成。社区是用户的松散组织，由具有相同爱好的，关注相同领域的用户组成，比如喜欢希区柯克的电影的人组成的社区，计算机领域研究人员构成的社区等。Web 中数据的高度异质性使个人数据的发现和管理面临重重困难，而 Web 上的这些具有相同爱好或关注相同领域的社区提供了一个新的发现和管理个人数据的途径。通过对社区相关信息的集成和分析，我们可以精确定位并抽取每个社区成员的个人信息，并进而为该成员构建 Web 上的个人数据空间，提供个人数据管理服务。

为了探索 Web 上的个人数据管理的方案，我们在学术社区信息管理的场景下提出了实体（Entity）和关联（Association）两个概念。

• 实体 — 构成 Web 上的社区的对象，包括人（People）和数据（Data）两个大类。

• 关联 — 社区中各个对象之间的相互关系，根据对象的不同可以划分为人与人之间的关系，人与数据之间的关系，数据与数据之间的关系三类。

C-DBLP 是 WAMDM 实验室为计算机领域研究人员这个社区构建的一个文献信息集成系统，在集成到的数据基础上，我们尝试为用户提供一些基于个人数据的服务。通过 Web 上期刊、会议的论文页面以及研究机构、研究人员的网站信息，我们成功地集成到了研究人员发表的文献情况，建立了一个包含 55000 余条文献，51000 多位作者的数据集合。在 C-DBLP 的数据集上，我们通过实体抽取（Entity Extraction）获得了作者（Author）、文献（Paper）、期刊（Journal）、会议（Conference）等实体，通过关联挖掘（Association Mining）得到了合作者关系（Co-author）、文献—作者关系等常见的实体之间的关联。在此基础上，我们将数据按各实体之间的关系进行了重构，形成了以作者为中心的信息组织形式，为用户提供基于作者的文献检索、文献关键字查询、研究者关系挖掘等服务。C-DBLP 系统的整体流程如图 3 所示。

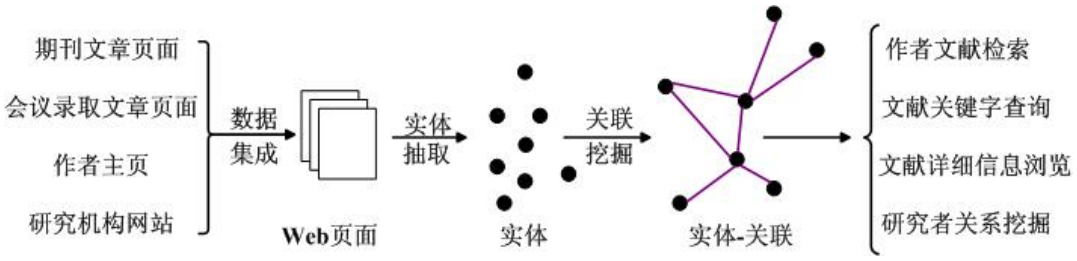

图 3 C-DBLP 系统整体流程

C-DBLP 成功地构建了一个基于文献信息的计算机领域研究人员的学术社区信息集成系统，我们把 Web 上该领域的文献信息集成到一起，通过实体抽取和关联挖掘过程获得了社区中成员及数据之间的相互关系，以社区成员为中心重构了来自 Web 上异质数据源的海量数据。我们将通过扩展数据源集成 Web 上更多的计算机领域研究者的相关数据，挖掘出更加丰富的实体和关联信息（如学生—导师关系，同事关系等）。通过 C-DBLP 系统的开发实践，我们认为，从社区的角度发现和管理 Web 上的个人数据具有一定的可行性。

Ⅲ 系统架构与实现

在 C-DBLP 系统的开发中，我们采用了模块化设计以保证系统的稳定性和可扩展性。根据业务流程实现的需要，我们将系统划分为三个模块，包括数据集成模块、数据处理模块和服务提供模块，分别完成对 Web 数据的集成、对原始数据的处理以及提供文献检索服务的功能。系统的整体架构如图 4 所示。

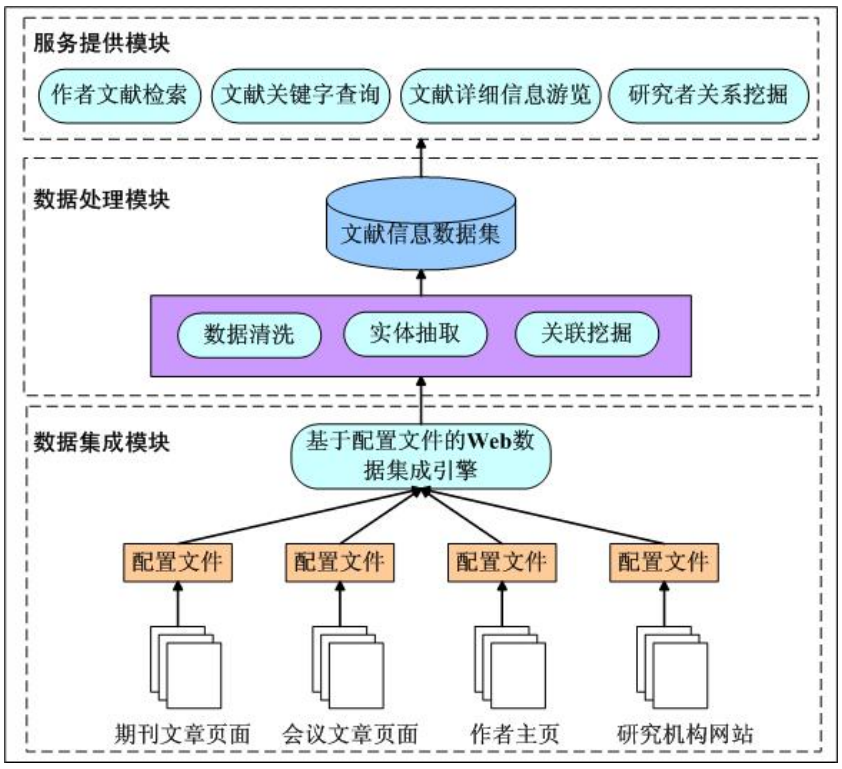

图 4 C-DBLP 系统架构

• 数据集成模块：

C-DBLP 的数据集成模块采用了 WAMDM 实验室在 Web 数据集成方面的研究成果—面向领域的 Deep Web 数据集成方案[1]，该方案中采用了基于配置文件的集成系统，用一个统一的基于配置文件的 Web 数据集成引擎，利用针对每个目标数据源网站的配置文件对 Web 上的数据进行集成。

面向领域的 Deep Web 数据集成方案是 WAMDM 实验室在 Deep Web 数据集成研究方面多年的研究成果的结晶，它既能够在某一个具体领域快速构造出一个大规模的数据集成系统，继而实现一个

垂直搜索引擎，也能够将各个不同的数据源的模式抽象出来构造 Mashup 应用。去年我们实验室利用该解决方案成功构建了工作信息集成引擎 Jobtong（http://www.jobtong.cn），而这次在 C-DBLP 系统的数据集成过程中的成功应用再次验证了这套方案的可行性。

• 数据处理模块：

由于 Web 上的数据的高度异质性，数据集成模块获得的原始数据并不能直接支持顶层应用的需求，因此我们引入了数据处理过程。数据处理包括三个部分，**数据清洗**功能从集成到的数据中去掉与文献内容不相关的信息（如期刊中的征文信息）以及重复的记录；**实体抽取**功能通过对原始数据的格式的分析抽取出每条信息中包含的实体，如文献作者（Author）、期刊名（Journal）等；**关联挖掘**功能则根据经过清洗的数据集以及抽取到的实体集合建立实体之间的关联关系，在目前的数据集中 C-DBLP 已建立了文献作者之间的合作关系（Co-author），对其它常见的关系的挖掘将随着数据集的扩充继续开展。数据处理部分还将文献信息以作者为中心进行了重新组织。

• 服务提供模块：

为了向用户提供以文献作者为中心的检索和浏览方式，我们在集成到的文献数据上建立了提供服务的网站（如图 5 所示）。C-DBLP 的网站向用户提供基于作者名的精确匹配检索以及基于作者、关键字、时间等信息的模糊查询。此外，系统还支持对文献详细信息的浏览以及对合作者信息的查询，力求将系统中所有有关文献作者的信息展示给用户。

图 5 C-DBLP 网站界面

Ⅳ 系统功能介绍

C-DBLP 系统利用面向领域的数据集成技术，从 Web 中集成计算机领域的部分权威中文期刊和学术会议论文的信息，向用户提供以文献作者为中心的检索服务。系统目前的文献收录情况如下：

◆已收录计算机领域的以下 12 本权威中文期刊的数据。

（1）软件学报 1990-2008 年

（2）计算机学报 1978-2008 年

（3）计算机研究与发展 1960-2008 年

（4）计算机工程 1975-2008 年
（5）中国图形图象学报 1996-2008 年
（6）中文信息学报 1986-2008 年
（7）计算机科学 1979-2008 年
（8）小型微型计算机系统 1980-2008 年
（9）计算机科学与探索 2007-2008 年
（10）中国科学 E 辑 1996-2008 年
（11）计算机辅助设计与图形学学报 1989-2008 年
（12）电子学报 1962-2008 年

◆已收录中国数据库学术会议（NDBC）2000 年-2008 年论文集的数据。

我们详细斟选了文献的数据来源，只收录国内计算机领域权威核心期刊和学术会议的数据。由于国内很多学术会议的论文集无法在 Web 上获得，所以目前 C-DBLP 中只收录到了中国数据库学术会议（NDBC）的最近几年的数据。

C-DBLP 尝试着通过对 Web 上的社区信息的集成和挖掘建立以社区成员为中心的个人数据管理方法，在计算机领域研究人员的社区中，我们通过对集成到的数据的分析处理初步建立了以文献作者为中心的个人学术信息空间，并为用户提供简单易用的检索功能：

• 以作者为中心的学术成果检索，为每位作者提供集成化的检索结果，展示该作者发表的中文论文情况，并展示该作者的合作作者情况。

• 提供基于作者名的精确匹配检索（Author Search）及基于作者名、论文题目、论文关键字、发表年份的模糊检索（Advanced Search）。

• 基于来源的文献浏览功能，系统支持对已收录文献按期刊出处和发表会议浏览。

• 浏览文献的详细信息，例如中英文题目、作者、摘要、关键字等。

我们将在 C-DBLP 中增加更多高质量的数据源扩充系统数据集，改进实体抽取和关联挖掘的算法，并引入新功能（如 Faceted Search 等），继续探索 Web 个人数据管理的方法，为用户提供更好的服务。

Ⅴ C-DBLP 中的开放问题

我们在设计和开发 C-DBLP 中文文献集成系统的过程中也遇到了一些棘手的问题，希望读者能够在这些问题上和我们一起探讨。

1．文献作者的重名问题

C-DBLP 系统目前的数据中已有 5 万余位作者，其中存在着一些作者的重名现象，很多情况下多位作者都拥有同一个名字，如“王伟”、“刘伟”等，这让我们在建立该作者与其它实体的关系时不能准确区分不同的作者。

关于重名区分（Name Disambiguation）问题，不少研究者都提出了自己的解决方案。基于概率模型的重名区分[2,3,4]将 Bayes 模型等概率分布引入作者的识别过程，但该方案的假设前提是作者满足一定的分布特征，这在 C-DBLP 的场景下并不适用。基于社会网络（Social Network）的重名区分方法[5,6]在社区信息管理中具有一定的可行性，但由于 C-DBLP 系统中集成到的数据并不足以体现各研究者之间的社会网络关系，该方案难以取得较好的效果。Yee Fan Tan 提出了使用搜索引擎返回的结果进行重名区分[7]，但 Web 上的中文名的搜索引擎返回的结果并不适用他提出的方法。

2．Faceted Search 的动态 facet 生成

随着新的数据源的不断引入和 C-DBLP 集成到的数据量的不断增加，用户在进行关键字查询时面对返回的大量结果会觉得无所适从，因此我们将在 C-DBLP 的数据集的基础上建立 Faceted Search 接口。如何在文献信息记录中抽取出能够代表该记录的 facet 是我们面临的最主要的问题。

Sanderson 提出了一种基于 subsumption 的从文本数据库抽取出代表文本内容的概念（concept）的方法[8]，但是该方法需要 $O(n^2)$的运算复杂度（n 为数据集中可能的概念的数量），这在 C-DBLP

海量数据的场景下代价太大。Wisam Dakka 对 Sanderson 的方法进行了改进[9]，他提出使用机器学习的方法抽取 facet 的方法，并通过假设关键字符合 Zipfian Distribution 将基于 subsumption 算法的时间复杂度降低到原来的 25%-30%，然而目前还没有能够支持文献信息处理的中文训练集。

3．学术社区成员间关联挖掘

Web 上的学术社区包含着成员及成员之间关系的丰富信息，根据之前提出的“关联”的定义，这些关系可划分为人—人关系、人—数据关系以及数据—数据关系。准确发现并挖掘出各成员之间的联系是社区信息管理需要实现的重要功能，而三类关系中又以人与人之间的关系最为重要也最难以获得。在 C-DBLP 系统中，我们通过对集成到的计算机领域的文献信息的关联挖掘获得了各成员之间的文章合作者关系（Co-author），但是这在计算机领域研究人员这个社区中是远远不够的。我们需要从 Web 上挖掘出更多成员之间的联系，如各文献作者之间的导师—学生关系、同事关系等。然而，由于这些信息较为特殊，我们在 Web 上还没有找到相关的数据源，因此我们希望读者能够对如何挖掘出社区成员间更多有价值的关联提供建议。

4．关联可视化处理

C-DBLP 致力于为用户提供最好的使用体验，我们希望能够把系统中的数据用更加大方美观的方式展现给用户。我们注意到 Web 上有一些很有特色的系统界面，如微软开发的人立方关系搜索引擎[10]，Searchme 公司推出的搜索引擎[11]等。我们希望在 C-DBLP 系统界面设计和搜索结果展示中引入更加丰富的可视化处理，使用户能够更好地享受检索过程，更方便地发现他们需要查找的数据。

Ⅵ 结束语

C-DBLP 是 WAMDM 利用 Web 数据集成研究成果建立的一个中文文献集成检索系统，我们在这个系统中进行了 Web 上的个人数据空间管理的探索和社区信息管理方法研究，提出了系统开发和功能扩展过程中遇到的一些开放问题。C-DBLP 既是一个公益性的文献集成项目，又是一个开展 Web 数据管理方法探索的平台。我们将在未来不断扩充系统的数据量，增强系统功能，为用户提供更加方便全面的服务。同时，我们将开发出系统编程接口，将该系统发展为一个简单易用的 Web 数据管理实验平台，希望能对读者的研究带来便利，也希望读者能够对 C-DBLP 中的问题提出更多更好的意见和建议，同我们一起把 C-DBLP 做得更加出色。

参考文献

[1] 王仲远，Jobtong：面向领域的 Deep Web 数据集成系统，WAMDM 2007 Annual Report: 57-61，January, 2008

[2] H. Han, H. Zha and C. Lee Giles. A Model-based K-means Algorithm for Name Disambiguation. In Proceedings of the 2nd International Semantic Web Conference, 2003.

[3] H. Han, W. Xu, H. Zha, C. Lee Giles. A Hierarchical Naïve Bayes Mixture Model for Name Disambiguation in Author Citations. In Proceedings of SAC’05, 2005

[4] D. Zhang, J. Tang, J. Li, K. Wang. A Constraint-based Probabilistic Framework for Name Disambiguation. In Proceedings of CIKM’07, 2007

[5] B. Malin. Unsupervised Name Disambiguation via Social Networks Similarity, In Proceedings of SIAM SDM Workshop on Link Analysis, Counterterrorism and Security, 2005

[6] B. Malin, E. Airoldi, K. Carley. A Network Analysis Model for Disambiguation of Names in Lists, Computational & Mathematical Organization Theory, 2005, Springer

[7] Yee Fan Tan, Min-Yen Kan and Dongwon Lee, Search Engine Driven Author Disambiguation. In Proceedings of JCDL’06, 2006

[8] M. Sanderson and W. B. Croft. Deriving Concept Hierarchies from Text. In Proceedings of SIGIR’99, 1999

[9] W. Dakka, P. G. Ipeirotis, K. R. Wood. Automatic Construction of Multifaceted Browsing Interfaces. In Proceedings of CIKM’05, 2005

[10]人立方搜索：http://renlifang.msra.cn

[11]SearchMe: http://www.searchme.com