对象级别的互联网垂直搜索

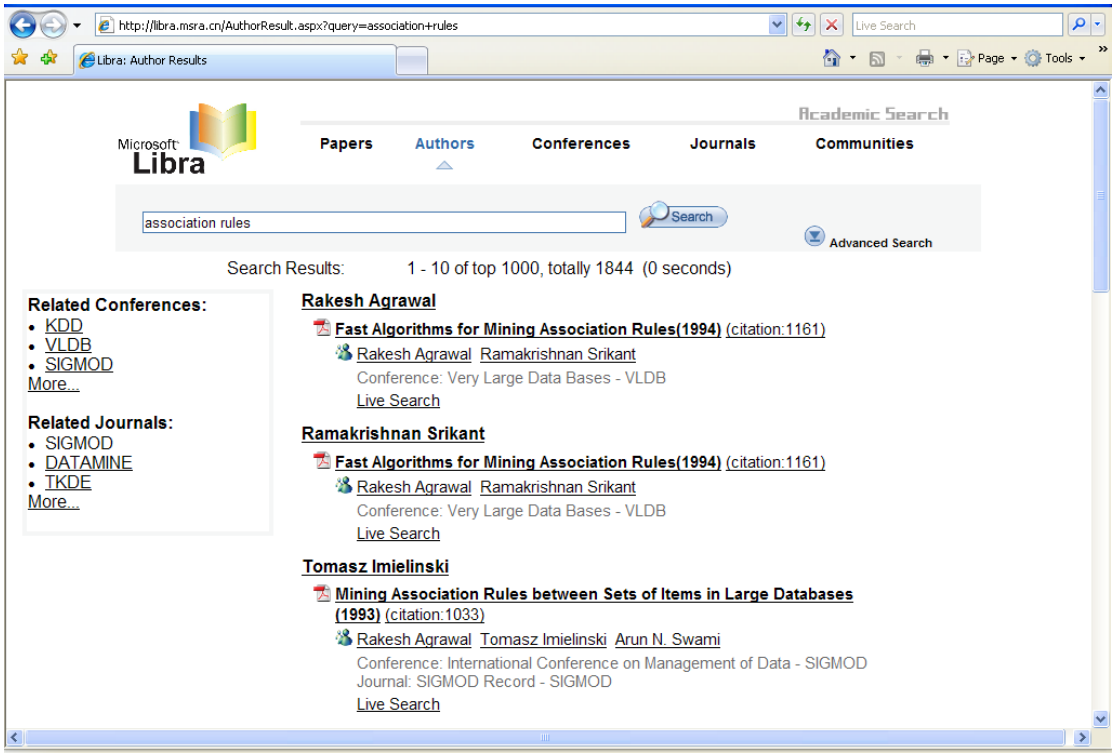
聂再清　　文继荣　　马维英
互联网搜索与挖掘组
微软亚洲研究院
{znie, jrwen, wyma}@microsoft.com

摘要

基于文档级别的排序和检索是信息检索领域几十年来的传统模式，目前大多数搜索引擎也采用了这种模式而把网页作为信息的基本单元。但是在网页和互联网数据库中存在着大量的关于真实世界中对象的结构化信息。在某些领域中，对象是一种更为有效的信息表述单元。目前我们正在探索一种新的模式，以实现特定领域中对象级别的互联网搜索，从而更好地满足用户的信息需求。这种对象级别的搜索模式由一系列相关技术构成，包括对象信息的发现和分类、对象抽取、对象继承、对象排序和检索等。在本文中，我们介绍对象级搜索引擎的系统体系结构和核心技术，以及两个具体的对象级搜索引擎：Libra 学术搜索和 Windows Live 商品搜索。

1．引文

现在主流的搜索引擎都把网页作为信息表示和检索的基本单元，这一思想是从传统的文本信息检索技术中继承而来的。在这一搜索模式下，关键词查询用于表达用户的信息需求，而查询结果是包含关键词并按相关性排序的网页列表。用户需要浏览列表中的网页以找到自己需要的信息。这一模式对于某些查询是有效的，例如查找主页、下载软件等。但是对于在某一特定领域内的复杂的信息需求，这种模式就不是很有效了。例如，我现在需要购买一个数码相机，我希望搜索引擎直接提供给我一个在指定价格范围内的数码相机列表，而不是网页列表。由于现有的搜索引擎是所谓的“通用引擎”，很多时候会反馈给我们大量与所搜索的内容毫无关系的结果。同时，对每种相机，我还想了解它的主要指标、用户评价、价格范围、出售商店等相关信息。可以想象，这样的搜索引擎能极大地方便用户的购物过程。



（图 1 注：Libra 学术搜索：当我们在 Libra 学术搜索引擎中输入某项技术的名称时，我们可以选择搜索其相关的文章、作者等所有相关内容。如果我们选择搜索相关作者，则搜索引擎会将这个领域中最著名的人物提供给我们，同时还会列出他的相关信息，例如地址、学术活动、所发表的论文以及个人网页等等。）

对象级别垂直搜索技术（Object-level Vertical search）的出现，正是为了解决这一问题。所谓“垂直”，是指这一搜索技术并不“包罗万象”，而是指向某一个特定的领域，例如学术、购物、求职等等，用户将在自己感兴趣的领域内进行搜索。而“对象”的概念指的则是搜索引擎在反馈搜索结果时，不再表现为一个个独立的页面，而是将各个页面中的关于真实世界中对象的结构化信息按照用户的需求集合成一个个完整的信息单元。例如，当我们搜索某个商品时，对象级垂直搜索技术将把这个商品的图片、参数、价格、商家的位置、用户的评价等等相关信息集合在一个页面上，让用户可以在第一时间得到最需要的内容。从这个意义上来说，垂直搜索技术与现有的页面搜索技术最大的一点不同就在于，它返回的是一个“不存在”的网页，是搜索引擎根据用户的需求而将各类相关信息集合到一起所生成的一个新页面。

对于现有的互联网搜索技术而言，要想将搜索到的全部内容都按照相同的对象集合起来，几乎是不可思议的。因为搜索到的内容太多了，涉及了多个不同的领域。而当我们把搜索的范围缩小到一个特定领域时，由于我们事先已经知道了这个领域中信息发布的大致规律，因此就可以较为容易地将各种信息归类。还是以购物为例，虽然现在世界上存在着数不胜数的电子购物网站，其编写语言、页面风格、排列模式等都各不相同，但是，购物类型的网站总会有一些相似的地方，例如他们都会有产品的图片、价格、介绍等。因此，对象级垂直搜索技术要做的就是从不同的网页中将它们提取出来。相对而言，这个工作就变得简单多了，是完全可行的。

为了让大家更具体地了解对象级别垂直搜索技术的高效性和可行性，我们开发了一个对象级别的互联网学术搜索引擎叫做 Libra（见图 1. http://libra.msra.cn）。与传统的页面级别的论文搜索引擎如 Google Scholar 和 CiteSeer 相比，Libra 能让用户方便地找到包括论文在内的学术对象信息，如科学家、会议、期刊、和学术社区。我们把网上的信息以对象为单元

进行信息抽取并集成起来形成一个学术对象信息仓库。这些结构化的对象信息可以用来回答用户的复杂查询如“在数据挖掘领域中近五年来最有影响力的科学家都是谁？”我们通过计算对象与查询的相关度和对象本身的重要性来对对象进行排序。这样既相关又重要的对象就会被排在查询结果的前面。

我们现在正在把对象级别垂直搜索的关键技术应用在 Window Live 商品搜索引擎（http://products.live.com) 的开发中。到目前为止，我们以把商品网页的自动分类和抽取技术完全应用到了 Window Live 商品搜索引擎测试版中。经过第一个月的试运行，我们已经自动的找到了 100，000 网上商家和 31，627，416 个网页，并从这些网页中抽取出了 800，000，000 个商品对象信息。我们相信，利用我们的互联网对象抽取和集成技术，Window Live 商品搜索引擎可以成为世界上最全的商品目录库。

２．系统体系结构

在图 2 中，我们给出对象级别的垂直搜索引擎的体系结构。首先我们需要一个网页爬虫来抓取某一个特定领域中的所有相关网页，并将这些网页按它们所包含的对象信息类型（如论文，作者主页，商品信息网页等）进行分类。对于每一个对象类型，我们训练一个对象信息抽取模型用来自动抽取这类网页中的对象信息。这些抽取出来的结构化的对象信息被集成到其对应的对象信息仓库。垂直搜索引擎就用这些对象信息仓库来回答用户的查询。当然我们可以基于这些结构化的对象信息做各种智能分析和挖掘工作，并用分析出来的知识来更准确地回答用户的高级查询。

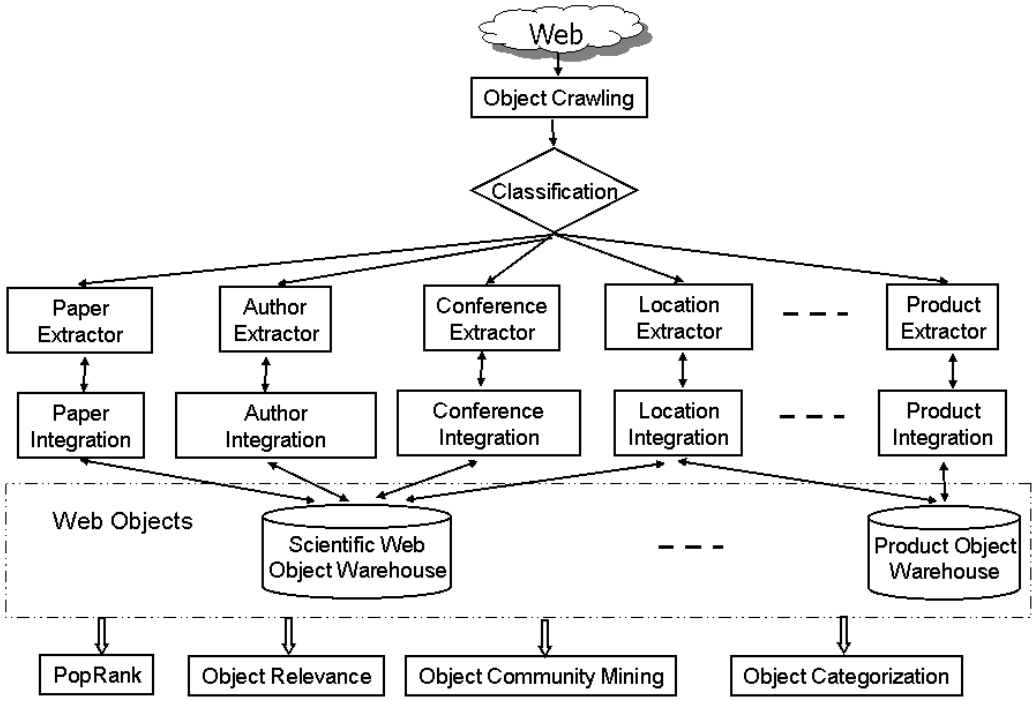


图 2．对象级别的垂直搜索引擎的体系结构

在具体地介绍对象级别垂直搜索的关键技术之前，我们先讨论对象级别垂直搜索的系统要

求：

1. 信息可靠性（Reliability）：高质量的结构化信息是让用户相信我们给出的查询结果的必要条件，否则用户可能更趋向于自己从网页中寻找答案。
2. 信息全面性（Completeness）：我们的对象信息必须很全面才能提供值得用户信赖的答案。
3. 排序的准确性（Ranking Accuracy）：在数以亿记的可能答案中，我们需要准确的对象排序算法以便让用户能够快速检索到所需的内容。
4. 规模可伸缩性（Scalability）：要包含在一个领域中所有的对象信息，我们需要存储所有在网上的和在各地离线数据库中的相关对象信息。这就意味着，对象级别垂直搜索技术必须拥有一个能够存储数以十万亿甚至百万亿记的超级数据库，并且还需要一套足够快的算法来快速检索到这些信息。

３．核心技术

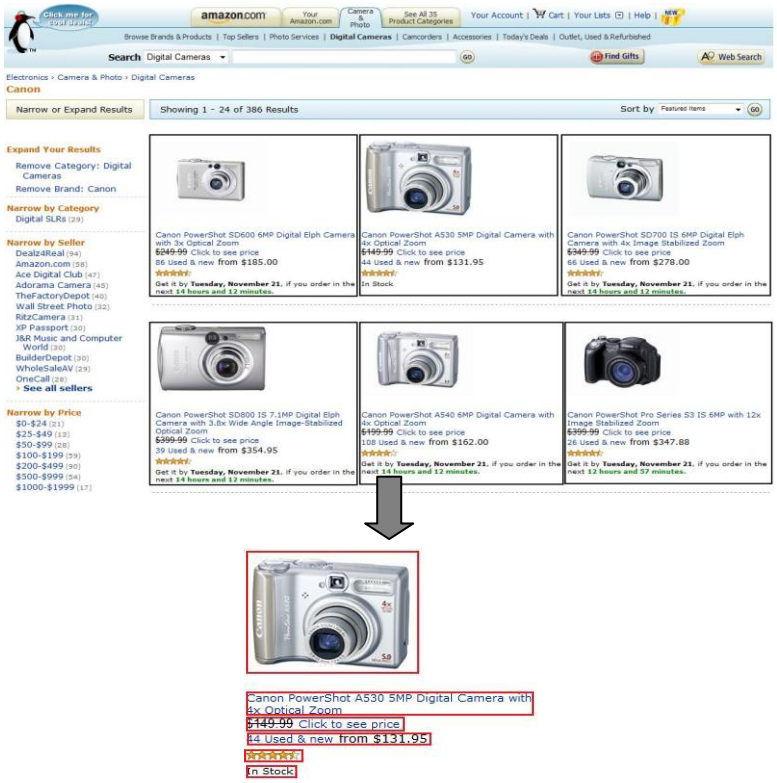
我们现在介绍对象级别垂直搜索中的关键技术：网页爬虫和分类、对象信息抽取、对象信息集成和对象排序。

网页爬虫和分类：我们知道，互联网上的信息浩如烟海，对象级别垂直搜索技术如果想为用户反馈如此精细的结果，就必须事先对所要查找的内容进行搜集和整理。我们扩展已有的聚焦爬虫（focused crawler）技术来在互联网上高效地抓取含有相关对象信息的网页。我们利用对象关系图来指导聚焦爬虫高效地找到相关网页。然后，通过一个“分类器”，将网页按照其所含的对象类别进行分类，例如学术论文网页、购物商品网页、作者个人主页等等。我们用支持向量机（Support Vector Machine）的方法对这些网页进行分类。因为分类器是和聚焦爬虫捆绑在一起的，分类速度要非常快。比如在我们的 Window Live 商品搜索中，分类器的速度小于０．１毫秒。为了达到快速准确地分类，我们利用很多启发式规则进行筛选。这些启发式规则往往能很快速地过滤掉很多无关的网页。比如“在英文的商品网页中必须出现‘$’符号”就能排除大量的非商品网页。

对象信息抽取：接下来，就需要对每类不同对象的网页进行对象信息提取了。所谓对象信息提取，就是按照这一类网页的特色，将其中的对象信息抽出来。例如购物商品网页，程序需要抽出来的就是产品的图片、价格、参数、评价等信息；而对于学术论文网页来说，则需要提取的就是作者名字、论文题目、文章内容、作者信息等等。

我们的对象信息抽取技术是一种基于条件随机场（Conditional Random Fields）的机器学习方法。首先，通过人工将部分网页进行标记，也就是将这个网页中各个关键部分标记出来，例如产品的图片、价格、参数、信息等。在积累了一定量的数据后，就开始让电脑自己对网页中的内容进行抽取。当然，在一开始，这个过程难免会出现错误，因此还需要人工进行纠正。不过，随着数据量的加大，其正确率会越来越高，直到基本上不再需要人工的参与也能够保持很高的准确率。

需要特别指出的是，对象级别垂直搜索技术中所采用的网页提取技术，与之前人们进行的类似研究相比，技术上最大的特色在于其使用了网页的视觉信息。我们认为，网页之所以比普通文本更加吸引人，就在于其有着丰富的视觉信息（见图３）。这些视觉信息让我们能够利用网页元素（HTML elements）之间的各种依赖关系（dependency）如二维依赖关系和层次依赖关系来进行更有效的对象信息抽取。具体的抽取技术请参见我们已发表的论文[5][6]。



（图 3 注：在提取网页信息时，对象信息抽取技术将充分考虑视觉对于人们观看网页时的重要性。）

对象信息集成：我们需要把从网上抽取出来的对象信息和现实世界中的真实对象一一对应起来，并把同一对象的所有信息（包括本地数据库中的已有信息）集成起来放入对象信息仓库中。这是一个典型的信息集成问题，包括两个子问题：

- **不一致的对象属性值的问题**：一个对象的某些属性有多个不一致的属性值。这些不一致可能是由不同的拼写方式或表现形式造成的。比如，“WWW”和“World Wide Web Conference”都可能指同一个国际会议的名字，一个是缩写而另一个是全称。
- **对象重名问题**：现实生活中经常出现不同对象的重名问题。比如在微软亚洲研究院就有好几个叫张磊的员工。我们有时会很难区分在一个论文中的作者是指的哪一个张磊。

虽然这两个子问题都很重要，对象重名问题更具有挑战性。因为解决重名问题往往需要利用比在对象信息仓库中已知对象属性值更多的信息。现在还没有很好的重名问题解决方案。大部分已有工作集中在利用已知文字属性来解决同名问题，近来也有一些工作开始利用对象之间的关系图来计算两个同名对象信息之间的关系强度。但是仅仅利用已知的文字属性和对象关系图很难准确地确定两个同名的对象信息是否指向同一个实体。因为在实际的应用中，经常出现信息不全的情况。在 Libra 学术搜索中我们经常发现论文引用连接关系的丢失。在这种情况下，计算出来的对象关系强度就不会反映真实的强度。

在开发 Libra 学术搜索和 Windows Live 商品搜索的过程中，我们发现含有同一对象的不同表象的上下文信息经常同时出现在同一个网页或网站中。比如，论文作者经常把他的所有论文放在他的个人主页上。我们把这种互联网上的不同表象的上下文信息的共同出现关系叫做互联网关系。我们利用上下文在互联网的共同出现情况来计算不同表象的互联网关系强度。我们用一种机器学习的方法来自动地计算互联网关系强度。当不同表象的上下文信息在同一网站的两个不同网页中出现时，我们通过考虑这两个网页的 URL 距离来计算互联网关系强度。具体的集成技术请参见我们已发表的论文[2]。

对象排序：当收到用户查询后，我们要计算对象与查询的相关性和对象本身的重要性。在计算对象相关性时我们要充分考虑在对象抽取时可能带来的错误和网上信息本身的不可靠性。传统的网页相关性计算不用考虑网页的错误是因为网页搜索是不考虑语义的。而对象搜索是有语义。因为我们会提供一个对象各种属性信息如商品的名称，图像，价格等。在我们的论文[1]中，我们提出一种对错误不敏感的对象相关性计算模型。在计算对象的重要性时，传统的网页重要性计算方法 Pagerank [4] 也已不再适用。因为在对象仓库中，我们能够看到各种不同的对象关系。比如在学术对象仓库中，有作者和论文的关系、论文和会议的关系、论文和论文的引用关系等。这些关系都能影响一篇论文的流行度（重要性）。因为一篇论文的重要性可能会受其作者、发表的会议和引用它的论文的重要性的影响。在[3]中，我们提出一个叫 PopRank 的对象重要性计算模型来考虑各种关系对对象重要性的影响。

从以上的描述我们不难看出，对象级别垂直搜索技术具有良好的通用性和学习性。对于一个对象级别的互联网垂直搜索引擎来说，这两点是至关重要的。前者可以保证这一引擎能够被迅速地应用于各个领域而不需要太大的调整，当然，一些必要的设置和资料的收集、整理还是不可或缺的；而后者则能够让这一引擎在进入新的领域时能够在最短的时间内投入使用，因为整个平台具有完善的自我学习能力，只需要前期的短暂调整，就可以自动而高速地学习新领域中的相关内容。

４．结论

在这篇论文中，我们提出一种新的搜索模式叫做对象级别的互联网垂直搜索。我们介绍了这种对象搜索模式的系统体系结构和核心技术。其中很多核心技术已经被成功的应用到了两个垂直搜索引擎当中：Libra 学术搜索和 Windows Live 商品搜索。我们相信这些技术具有很强的通用性，可以广泛地应用到到多数的垂直搜索引擎中如黄页搜索、博客搜索、人物搜索、工作职位搜索和饭馆搜索等。

５．参考文献

[1] Zaiqing Nie, Yunxiao Ma, Shuming Shi, Ji-Rong Wen, Wei-Ying Ma. Web Object Retrieval. To appear *in the Proceedings of the 16th international World Wide Web conference (**WWW 2007**)*.

[2] Zaiqing Nie, Ji-Rong Wen, Wei-Ying Ma. Object-Level Vertical Search. In the *Third Biennial Conference on Innovative Data Systems Research* (CIDR 2007).

[3] Zaiqing Nie, Yuanzhi Zhang, Ji-Rong Wen, and Wei-Ying Ma. Object-level Ranking: Bringing Order to web Objects. In Proc. WWW, 2005.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Libraries, 1998.

[5] Jun Zhu, Zaiqing Nie, Ji-Rong Wen, Bo Zhang, and Wei-Ying Ma. 2D Conditional Random Fields for web Information Extraction. In *Proc. of ICML*, 2005.

[6] Jun Zhu, Zaiqing Nie, Ji-Rong Wen, Bo Zhang, and Wei-Ying Ma. Simultaneous Record Detection and Attribute Labeling in web Data Extraction. In *Proc. of SIGKDD*, 2006