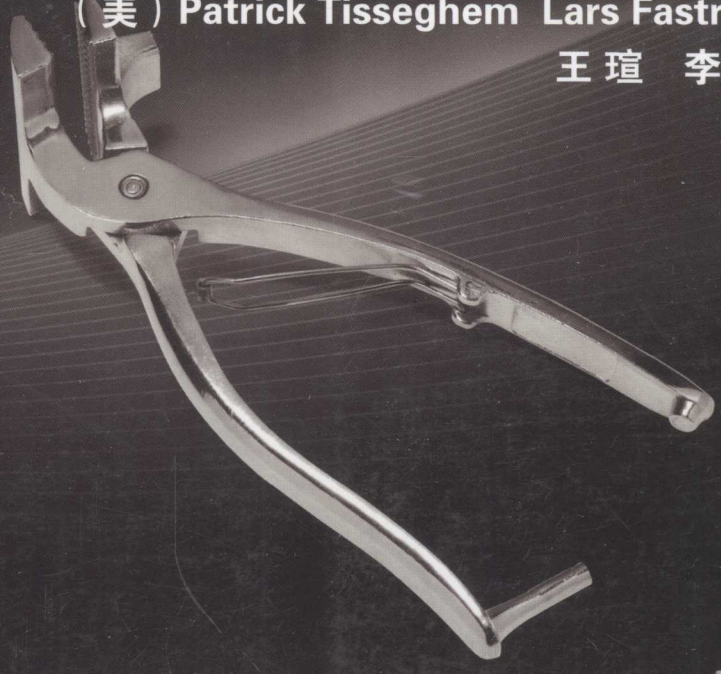
Microsoft

微软技术丛书

MVP
Microsoft®
Most Valuable
Professional

深入索引和搜索引擎

（美）Patrick Tisseghem Lars Fastrup 著
王瑄 李燕 译



- 涵括SharePoint Server 2007管理和开发两大主题
- 深入探讨搜索定制和优化技巧

清华大学出版社

微软技术丛书

MVP
Microsoft®
Most Valuable
Professional

深入索引和搜索引擎

（美）Patrick Tisseghem Lars Fastrup 著
王瑄 李燕 译

清华大学出版社
北京

《微软技术丛书》出版前言

在黄昏里希冀皓月与繁星
在深夜希冀着黎明
在炎夏希冀凉秋
在严冬又希冀新春
这不断的希冀啊，
使我感触到世界的存在，
带给我多量的生命的力。
这样，
我才能跨过——
这黎明黄昏，黄昏黎明，春夏秋冬，秋冬春夏的茫茫的时间的大海啊。

——艾青

时间在流逝，技术也在迅猛发展。在希冀中，微软的.NET 战略早已经变成现实，带来全新、快速而敏捷的企业计算能力，也给软件开发商和软件开发人员提供了支持未来计算的高效 Web 服务开发工具。在希冀中，我们欣喜地看到，微软的每一个技术创新，都对中国开发人员产生巨大的推动作用，使得越来越多的人加入微软开发阵营。

微软出版社为了配合 Visual Studio 的推广和普及，邀请项目开发组的核心开发人员和计算机图书专业作家精心编写了微软 IT Pro 系列图书。该丛书自面市以来，在美国图书销量排行榜上一直高居前列，颇受读者好评，成为程序开发人员和网络开发人员了解微软技术的权威工具书。随着新的开发平台的发布，该系列得以大幅度扩充，在美国及欧洲图书市场广受好评。

从 2002 年开始，清华大学出版社为了满足中国广大程序开发人员、网络开发人员以及计算机用户学习最新技术的渴望，在微软出版社的配合下，先后推出了《微软.NET 程序员系列》和《微软.NET 程序设计系列》。这两套书阵容庞大，几乎涵盖.NET 技术及其应用的各个方面；也正因为如此，翻译和编辑加工的工作量也大得惊人。但为了保持国外优秀技术图书的魅力，同时使读者领会新技术的真谛，本丛书的翻译和编辑都是经过严格筛选的、具有很高的翻译水平或丰富编辑经验的技术人员。同时，我们还聘请微软公司相关产品组的技术专家审读每一本书，确保在技术上准确无误。

2005 年，随着微软新的开发平台的推出，我们将原有的两套丛书整合为《微软技术丛书》。这套丛书针对不同层次的读者，分为 5 个子系列：从入门到精通、技术内幕、高级编程、精通&宝典和认证考试教材。各系列特色如下：

★ **从入门到精通**

- 适合新手程序员的实用教程
- 侧重于基础技术和特征

- 提供范例文件

★ **技术内幕**

- 权威、必备的参考大全
- 包含丰富、实用的范例代码
- 帮助读者熟练掌握微软技术

★ **高级编程**

- 侧重于高级特性、技术和解决问题
- 包含丰富、适用性强的范例代码
- 帮助读者精通微软技术

★ **精通&宝典**

- 着重剖析应用技巧，以帮助提高工作效率
- 主题包括办公应用和开发工具

★ **认证考试教材**

- 提供完整的 Ebook(英文版)
- 提供实际场景、案例分析和故障诊断实验
- 完全根据考试要求来阐述每一个知识点

这套丛书延续以前严谨的编校风格，一切以保证图书内容和技术质量为核心，付出了大量心血。相信整合后的这套丛书必然会帮助程序开发人员、网络开发人员以及具有一定编程基础的中高级读者，快速、全面地掌握微软技术，为将来的技术生涯奠定扎实的基础，使之成为中国软件产业的栋梁！

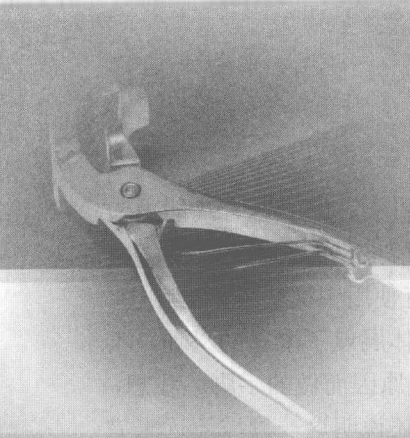
为增强本书的可读性，便于读者迅速定位关键术语的原文和快速根据索引来定位知识点(概念、函数等)的详细介绍，有些经典图书中在相应位置标注了原书页码(在当前行末尾用粗体方括号【】或椭圆形底纹 表示)，并在书后附上原书索引，以期能对大家提供更多的帮助。已经采用这一体系设计的图书有《Windows 核心编程(第 5 版)》、《Visual C# 2008 从入门到精通》、《ASP.NET 3.5 核心编程》、《Visual C# 2008 核心编程》和《精通 Windows 3D 图形编程》。

在此，感谢参与本丛书的翻译和审校人员，感谢他们付出的心血和时间。他们来自培训和实践前沿，具有深厚的技术底蕴和文化素养，善于用浅显易懂的语言阐述晦涩难懂的技术细节。同时也要感谢这一年来时刻关注这套书的读者朋友们。他们热心地提出自己的意见和建议，感谢他们的宽容和善意关爱。我们将和大家一样，时刻关注微软技术发展的最新动态，时刻保持自己的技术动力！

亲爱的读者朋友，期待着您把每一次看书的机会，都当成增进知识的时候。这个过程，绝对不是浅尝辄止，更非自认把书看过一两遍就可以了。深度的阅读是尽可能地把书本的知识转换为自己熟悉的，甚至读到自己内心的深处。同时，也请把您对这套书的感受告诉我们，我们期待着和您分享，联系信箱 coo@netease.com。

尽管我们注入大量心血，但疏忽纰漏之处在所难免，恳请读者朋友提出建议和批评。本丛书在创作、翻译和编辑过程中得到了微软(中国)公司的大力支持。本丛书能够顺利出版，更是倾注了无数幕后人员的汗水和心力。在此，对他们的辛勤劳动一并表示衷心感谢！

清华大学出版社



译 者 序

穿越无数个蛮荒蒙昧的黑夜，人类借由一次次科学技术的演进不断探索着自身理想的生存方式，从这样的意义上讲，互联网毫无疑问是一场伟大的革命，虽然我们无法猜测创建者的初衷，但是我们可以看到它的实际作用：人类再一次尝试为所有人提供公平的机会！

透过互联网，人们可以迅速获取各种知识而成为“全能”的专家，可以以最低的成本展示自己的才华，可以不戴任何面具说出自己真实的想法，可以以理性的方式说服别人或者被别人说服，更重要的是，凭借搜索引擎的帮助，我们可以从容地在各种事物之中做出选择。

在信息无比浩瀚的互联网中，一款好用的搜索引擎可以节省使用者的不少时间和精力。当下，Google 无疑是这一领域的佼佼者，同时，百度在国内也深受欢迎。两家搜索引擎公司都已经在站稳各自搜索市场的基础上开始在其它领域进行业务拓展，以期为用户提供更多的服务。

在互联网发展的过程中微软给人的感觉是后知后觉，在现在的互联网搜索领域，它也绝对谈不上业界领先，不过考虑到其实力和决心，以及后来居上的“前科”，我们却不能小看它，尤其是在我们了解了微软制订的占领市场的策略之后。

微软应该是很清楚自己目前的优势和劣势，虽然还不具备在“正面战场”上与 Google 竞争的实力，但是凭借其在操作系统和办公软件市场占有率的绝对优势，微软将突破的重点放在了“企业级”搜索市场，换言之，微软希望先让人们熟悉工作环境中的微软搜索引擎，进而乐于使用互联网领域的微软搜索引擎。

于是，这一重任自然就由一直作为企业协作平台的 SharePoint 系列产品肩负起来了。我觉得微软的这个主意相当棒，Google 虽好，却不能帮助企业解决一切问题。每个企业情况都不尽相同，这些企业通常期望能对企业内部使用的搜索引擎进行更多的控制，这些控制包括更加符合自身企业文化的用户使用界面、与正在使用的办公软件的整合、指定的搜索结果来源、搜索结果的排序以及企业经理最希望用户看到的搜索结果……如果企业已经使用微软的 SharePoint 构建自己的协作平台和门户站点，那么企业中的信息管理人员一定乐于见到 SharePoint 中整合了一个可以解决上述所有问题的强大搜索引擎，只要按照微软一贯坚持的并且已经为人们所熟悉的操作方式，信息管理人员们就可以很方便地对该搜索引擎进行配置和改造，以满足各自企业的使用需求。在此过程中，我们还将惊喜地发现：只要稍稍掌握一些编程技巧，我们还可以让使用者更加满意。

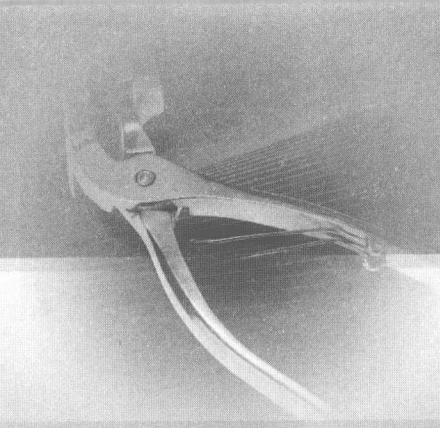
本书旨在指导读者完成上述工作。作者 Patrick Tisseghem 和 Lars Fastrup 拥有丰富的 SharePoint 相关使用、开发和教学经验，这使得本书内容的权威性不容置疑。在翻译过程中，

我们力图保持原文的这种权威性，当然，我们也努力让两位计算机学者的表达方式能更为中国读者所接受。水平所限，错误在所难免，烦请广大读者批评指正。需要指出，由于数据方面的限制，我们在译稿中保留了部分原书使用的图片，而没有使用中文环境下的截图。

有趣的是，撰写这篇序言的时候，人们突然又对搜索引擎集体展现了浓厚的兴趣，不过这一次是因为媒体曝光了 Google 和百度在搜索排名上因为利益原因而做的手脚。我想，这不光说明互联网与现实社会的联系已经到了多么紧密的程度，还说明这一领域还远未完善，前路漫漫，我们要做的还有很多，无论是技术还是其他……

感谢参与翻译工作的众位好友：田小梅、朱艳、王大平、金立年、王谨承、陆璠、周龙川、蔡亚忠。

感谢清华出版社的文开琪小姐，您的热情和严谨给予我们莫大鼓舞。



致　谢

在需要每天按时上下班的情况下完成一本书的写作，无论对个人，还是对整个家庭，都是一项非常有挑战性的工作。这就是我希望首先感谢我的家人的原因，感谢我的妻子 Linda 和我两个可爱的女儿，Laura 和 Anahi，感谢她们在过去 8 个月里给我的支持。

在 2003 年夏天我开办的第一个 SharePoint 2003 开发者培训课上，我认识了 Lars，他当时是我的学生之一。从那时候开始，我们便成了非常好的朋友。由于在该领域拥有出色的经验，他决定与我一起写作本书。我非常感谢他付出的努力。如果没有他，本书就会变成一本只适合开发者阅读的书籍，而 Lars 确保了书中有大量供管理人员参考的内容。

我还必须感谢 Karine 和整个 U2U 团队的支持，感谢微软出版社，感谢 Lori、Steve、Jennifer、Randall 和 Rosemary 等诸位编辑，感谢 Mike Fitzmaurice 的不懈支持(甚至是在 Antartic)，感谢我的学生们，过去两年我在不同国家所开办的培训课上，他们给予了我很多启发。

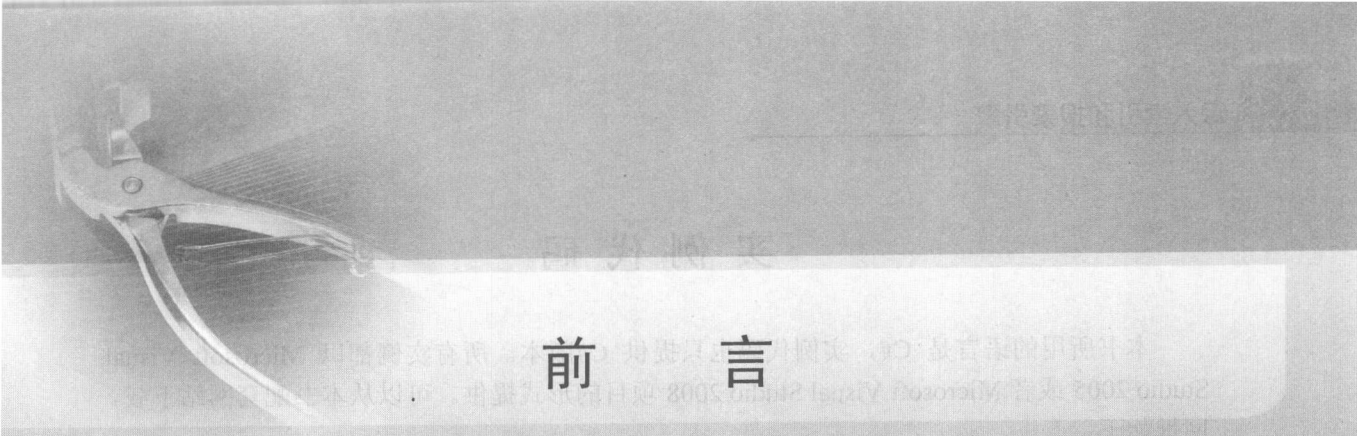
——Patrick Tisseghem

这是我写的第一本关于 SharePoint 的书籍。首先，我需要感谢我的合作伙伴 Patrick Tisseghem，感谢他给我这样一个合作写作的机会，虽然为了达到他的工作要求我付出了很多努力，但与他一起工作是非常有挑战性并且非常有价值的经历。他随和的性格与我相得益彰，我非常喜欢与他一起共事。我特别怀念在丹麦我的住处我们进行的第一次讨论会议，怀念我俩共同喜爱的单麦芽威士忌和那可爱的夜晚。

感谢微软出版社，感谢你们信任我并接纳我成为本书的作者。我希望着重感谢我的主要合作伙伴，他们是 Lori Merrick、RoseMary Caperton、Steve Sagman、Randall Galloway 和 Jennifer Harris，感谢他们的耐心以及对我所写内容卓有成效的审校工作。他们对于最后成书所做的莫大贡献令我印象深刻。

最后，我还希望感谢微软及 SharePoint 搜索功能项目组，感谢他们开发了这样一个拥有巨大价值的产品并推向市场。这款产品在我的专业职业生涯中扮演着重要的角色，我利用它建立了一个出色的专业化网络和一整套技术技能知识库。我简直非常渴望知道下一个版本会呈现哪些新花样。

——Lars Fastrup



前　言

十年来，微软对企业级搜索领域进行大幅投入，并有日益增加之势，这是我们有目共睹的。在写作本书的同时，我们已经在 Windows SharePoint Services 3.0 和 Microsoft Office SharePoint Server 2007 中实现了对搜索功能的支持，并且，新发布的 Microsoft Search Server 2008 和 Community(社区)工具集还拓展和完善了搜索的功能架构及用户搜索体验。它们是致力于企业级搜索领域的微软公司的一笔财富，它们能够帮助微软公司在竞争日益激烈的市场中拼杀，销售扩展组件提升管理人员、开发人员及客户的使用体验，以及为他们提供咨询服务。

2007 年夏天，我们(Lars Fastrup 和 Patrick Tisseghem)决定合作写作本书。从那时到现在，我们始终确信，写一本涵盖在一个组织中部署 Microsoft Office SharePoint Server 2007 的书正当其时。Lars 的大部分职业生涯都献给了一款名为 Ontolica Search 的极为成功的第三方产品。Patrick 是《SharePoint Server 2007 实用宝典》(微软出版社 2007 年出版)一书的作者，他负责 SharePoint 的开发培训，并且他对使用定制解决方案组件来定制及扩展搜索架构课题有着极大的热情。

目 标 读 者

本书涵盖管理人员和开发人员所关心的广泛主题。正如阅读本书时你将发现的那样，很多时候，管理人员和开发人员之间的职责并不是那么泾渭分明的。为了很好地完成工作，管理人员和开发人员都必须加深对对方工作的了解。

当然，本书参考了大量 Windows SharePoint Services 3.0 和 Microsoft Office SharePoint Server 2007 的 Software Developer Kit(SDK)中的内容。在阅读本书的同时，使用这些资源是非常有帮助的。可以在微软开发者网络(MSDN)中找到这些资源，网址为 http://msdn.microsoft.com/en-us/library/bb931736.aspx，或访问 http://www.microsoft.com/downloads/details.aspx?familyid=6d94e307-67d9-41ac-b2d6-0074d6286fa9，下载完整的 SharePoint Server 2007 SDK 文件。管理人员可以在 http://technet.microsoft.com/en-us/library/cc263630.aspx 得到微软 Technet 站点更多的背景资料。

实 例 代 码

本书所用的语言是 C#，实例代码也只提供 C#版本。所有实例都以 Microsoft Visual Studio 2005 或者 Microsoft Visual Studio 2008 项目的形式提供。可以从本书配套网站下载，网址如下：

http://www.microsoft.com/mspress/companion/9780735625358/

微软出版社还在以下网站提供对图书的售后服务和配套内容：

http://www.microsoft.com/learning/support/books/

问题与意见

如果对本书及其配套内容有任何意见和建议，或者在浏览上述网站后依然心存疑问，请通过电子邮件的方式告知我们：

mspinput@microsoft.com

注意：该邮件地址并不提供对微软软件产品的支持。

目　录

第 1 章　SharePoint 2007 企业搜索功能简介

学习完本章内容之后，您将能够：

- 描述微软在企业搜索领域中所扮演的角色
- 区分各种微软企业搜索产品所支持的搜索功能
- 知道 Search Server 2008 是如何支持联合搜索功能的
- 较为深入地理解 Office SharePoint 产品的索引和搜索架构
- 描述本书所包括的面向管理人员和开发人员的主题

1.1　搜索的重要性和微软所扮演的角色

“知识有两种，一种是知识本体，另一种是我们获取知识的方法。”时至今日，Samuel Johnson 在 18 世纪的思考显然仍不落伍。以数字技术存储的信息数量极为庞大，并且日益增加。如今，如果没有软件工具帮助我们适时地找到有用的信息，人类几乎无法全面而深入地认知我们所生活的这个世界。不仅在个人的工作中我们能够感受到这一点，在一个由不同人员组成的组织中，这种感受更为深刻。组织机构中的信息工作人员迫切需要使用专业化软件从存储在不同位置的庞杂数据中查找到组织合作信息。

过去十年来，微软对于互联网、桌面和组织内联网(或称企业内联网)这三大搜索领域的相关搜索技术进行了大量的投入。微软通过 Live Search(http://www.live.com)实现互联网上的搜索，见图 1-1。尽管 Live Search 面对的是与 Google 和 Yahoo 这样的搜索行业巨头的艰难竞争，但是最近的报表显示了其颇具成长性的市场占有率，这得益于 Live Search 最强大的特性：与其他微软在线服务的整合性。

在桌面搜索领域，凭借在最新终端操作系统中所提供的索引和搜索功能，微软处于强势地位。图 1-2 显示了在微软 Windows Vista 操作系统中提供的搜索功能。

许多用户认为，Windows Vista“开始”菜单(如图 1-2 所示)和 Windows 资源管理器(如图 1-3 所示)的搜索功能是该操作系统最棒的新特性之一。所有存储在笔记本电脑或者台式机里的数据都以一种井然有序的方式在后台编制有索引，且当需要这些数据的时候，用户根本不用操心它们的存储位置，很快就能找到它们。流行软件产品(如 Outlook 2007)中对搜索功能的支持明显为信息工作人员提供了很大的方便，假如需要从邮箱中保存的大量邮件中筛选出需要的信息，他们完全可以借助于搜索功能，方便地完成该项工作。

图 1-1 微软的 Live Search

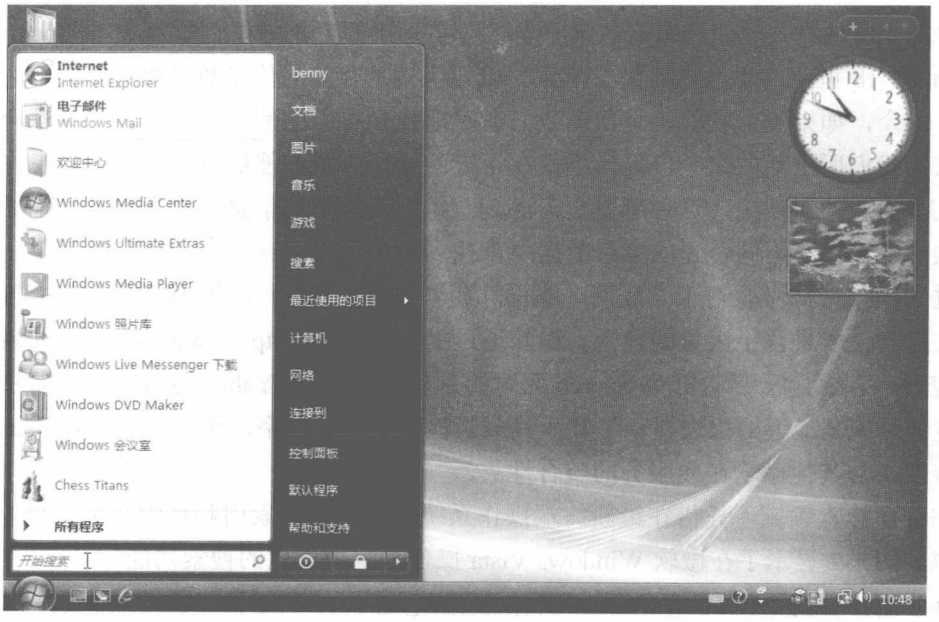


图 1-2 Windows Vista 的“开始”菜单中的桌面搜索功能

互联网和桌面搜索领域并不是本书讨论的主题。我们将把注意力集中在微软对企业内联网中所存储数据提供的搜索和索引支持上。这个领域的搜索支持出现在十年以前，当时微软有一款名为 Site Server 3.0 的产品，它使用 Internet 信息服务器(IIS)早期版本中包含的 Windows 索引服务，并且它还是微软 SharePoint 系列产品和技术的前身。企业内联网的搜索最初只支持 SQL Server 此类关系型数据库中存储的数据，但现在已经扩展到支持其他多种数据存储系统中存储的数据。

图 1-3　Windows Vista 中 Windows 资源管理器所提供的桌面搜索功能

虽然最初这几个领域的搜索技术平行发展，但时至今日，微软正努力将所有与搜索有关的技术整合到一项核心的搜索技术中去，使其能够方便地融入公司的大部分产品中，且可以根据实际情况进行定制以满足不同的需求。显然，公司的工作人员们不应该只被局限于对存储在公司内部数据库系统或共享文件夹中存储的数据进行搜索并得到数据。信息工作人员希望能够从各处得到信息技术支持：外部世界、本地文件系统和公司内部存储的数据(无论它们以何种形式存储)。这种将多个领域的搜索整合在一起的做法就是我们通常所说的企业搜索。

更多信息：要想进一步了解企业搜索，请访问 http://www.microsoft.com/enterprisesearch。

企业搜索并不只是对信息的索引和搜索，微软的想法是进一步把搜索结果变成动态的。简而言之，微软有三个设计目标：改善终端用户的搜索体验；在搜索结果中包含人们的专业经验；确保能够尽快以最合理的顺序得到搜索结果。而这一切都应该在管理人员的掌控之中。

1.1.1　用户搜索体验

对使用微软桌面和服务器产品的用户来说，进行信息搜索就像家常便饭一样平常。只要身处像 Outlook、Word 或者 Windows SharePoint Services 文档工作空间这样的工作环境，用户就可以输入搜索请求，让这些查询在后台执行，然后经由相同的操作环境将搜索结果返回给用户，使其得以及时处理。搜索结果中可能包含从不同来源得到的信息，这些来源

可能是互联网上类似于 blog(博客)和 Wiki(维基)这样的公共空间，可能是公司内网中存储办公系统相关文档的共享目录，也可能是连接在某个员工笔记本电脑上的外接硬盘。换句话说，用户可以在一个地方实现对多个位置信息的搜索。

搜索所属的上下文同样重要且应该在最终搜索结果的显示中有所体现。例如对于搜索同一个问题，北美地区的售货员和欧洲地区的售货员不应该看到相同的搜索结果。许多公司已经捕获并保存了一些上下文相关的信息，如工作环境和员工工作职责信息，并将它们以某种形式存储起来供人们使用。搜索引擎应该能够将这些信息提取出来并按照规则对它们进行整理和排序。而且，用户应该只能够查看查询结果中按照访问权限列表(ACL)规定允许他们查看的信息，该列表由存储这些信息的拥有者定义。在有些情况下，保密信息的访问权限由信息本身的独立区块来定义。通常情况下，对搜索结果的整理是必需的，而且对用户和负责管理搜索架构的管理员来说都应该是透明的。

仅仅简单地将搜索结果呈现给用户是不够的。搜索结果应该是动态的，而其动态性来源于搜索所属的上下文。对同一个问题的搜索，用户从 Outlook 中得到的搜索结果应该不同于从使用 Office SharePoint Server 2007 建立的门户网站中得到的搜索结果。当然，并不是只有搜索所属的宿主应用程序上下文会对搜索结果产生重要影响，发出搜索请求的用户和其他所有用户配置信息都会使整个搜索结果或者单一的搜索结果发生改变。

1.1.2 企业中的员工

对于企业来说，除了内网中各种设备存储的信息以外，还有另一种重要的资源：人。人有技能，人懂知识，人知道请求能人的帮助，人可以与别人分享自己所拥有的资源。作为企业，非常有必要将所有这些专业资源收集整理起来，并允许员工进行查询。在当今的互联网中，为人们提供社交、协作和分享等服务的网站(如 Del.icio.us、Facebook 和 LinkedIn)方兴未艾，另外还有其他许多与其类似的网站。而且，此类新网站每天都在大量涌现。我们希望通过访问这些网站来更多地了解别人，而作为访问者的我们正是这些网站的核心价值所在。如今，越来越多的公司在考虑如何利用软件在本公司内部建立这种社交网络。

应该说，人员信息的收集和管理并不总是轻松的事情。从事该工作的软件系统从存储人力资源信息的典型 LOB(line-of-business)系统到松散合作网站以及个人或专业博客，不一而足。以博客为例，它既可以仅限于公司内部交流使用，也可以作为公司对外形象宣传的一种手段。无论构建网站的目的是什么，都应该确保这些网站上的信息及其后台系统中的数据被编制了索引以便于查询。作为负责索引实际内容的索引架构的重要组成部分，爬网程序应该能够提取其所能找到的所有关于人员的信息并将它们提供给搜索引擎使用。当然，对信息的查找必须遵守法律条款，如个人数据的隐私保护条款。

搜索引擎应该能够在显示这些搜索结果的同时告诉用户可以对这些结果进行哪些操作。图 1-4 呈现了这样一个示例：对人员的查询列出了查询结果条目，并且列出了用户可以对其中某个查询结果执行的操作。