网络环境下的中文搜索引擎技术分析

申玉江 丁伟
（东南大学计算机科学与工程系，南京 210096）

【摘要】本文从信息检索和体系结构这两个角度对网络环境下的搜索引擎技术进行了一些分析研究。在信息检索方面，介绍了几种适用于中文的信息检索方法,并进行了一些分析比较。在体系结构方面,基于分布式处理的思路,提出了搜索引擎的分布式结构和相应的服务器端搜集代理的概念。

【关键词】搜索引擎 中文信息检索 计算机网络 分布式体系结构

1．引言

网上信息的分布性和随意性，使搜索引擎技术得到迅速的发展。搜索引擎是一个涉及信息检索、语言编码、计算机网络等多个领域和学科技术的综合性软件系统。现有大部分搜索引擎都是集中式的，从工作机制上看，主要是由两个相对独立的部分所组成：

1） **信息的搜集**

一般的搜索引擎都采取 ROBOT 的方法搜索各网页，尽可能多的将远处信息搜集到本地来，进行相应处理以形成本地信息源。

2） **服务的提供**

它是指采取某种信息检索方法从本地信息源提取合适的信息提供给用户。

在这个相对简单的工作机制中，信息检索技术是核心，不同的检索方法造成了各搜索引擎搜索效果和工作效率的千差万别。

应该承认，信息检索技术经过多年的研究和发展已经相对成熟，因此这种以信息检索技术为核心的集中式搜索引擎技术的发展也将为此受到很大的限制。毫无疑问，从其它研究方向的角度来探讨搜索引擎问题，并将相应的技术思路带入这个领域，可以为其发展赋予新的活力。

本文将在下面的篇幅里围绕信息检索和搜索引擎的体系结构以及相关的问题进行有关的探讨。

2．中文搜索引擎

由于起步较晚，与西文搜索引擎相比，目前网上现有的搜索引擎的搜索结果精确度相对较差，内容也比较单一。除了一些技术和非技术的因素外，汉语的下列基本特点使得它不如西文那样便于计算机处理。

1） 汉语是一种内涵语言，具有很强的概括抽象性。

2） 汉语的虚词作用独特，兼具逻辑连接词、时态算子、模态算子等功能，在时态、语气、逻辑关系等方面有独到的表达能力。

3） 主谓语难以确定，句法多变。

4） 汉语的词不是自然出现的单位，而是理解的单位。

所以在用计算机对汉语进行分词和语义诠释时，就不如以空格分隔的、句法相对确定的字母体西文那样有效，从而导致了中文搜索引擎的搜索效果不尽人意。解决这一问题的根本途径是分析研究针对中文的信息检索技术。当然除了搜索结果方面的问题外，还有运行效率方面的问题，如服务器负担过重、网络带宽浪费严重、重复建设、难以适应网络规模的日益扩大等。这是各中文搜索引擎普遍面临的问题。

3．中文信息检索技术

3．1 全文检索技术

这种技术是将搜集来的信息全文保留在本地数据库中，其基本检索方法是根据文档与用户查询串的相关程度打分，然后按分数由大到小排列。相关程度由查询串在文档中出现的频率，以及查询串在数据库中的整体比重有关。这种方法简单易行，但由于需要保存信息全文，所以需要海量的存储，而且它的全文匹配的串频统计方法会造成较大的查询响应时间，因此它难以适应网络规模的日益扩大，更重要的，它没有对查询串的切分过程，当查询串较长时，效果会很差。

3．2 索引检索技术

与全文检索技术不同，这种方法将搜集来的 WEB 页面进行概括或索引，得出关键性的、能体现全文大意的摘要信息入库，从而使存储量大大降低（摘要的生成方式与搜索引擎的工作方式密切相关）。在进行检索时，则利用查询串与这些摘要信息的关系来决定向用户提供什么样的信息。

目前应用于西文索引检索并取得良好效果的方法有很多，如 RSV 法、聚类法、相关矩阵法等，但它们却不能直接应用于中文，因为这些方法都必须从查询串和文档中提取 feature（所谓 feature，就是能在语义上代表文档或查询串的单词或短语序列）。这一工作对于西文来说比较简单。因为西文文档的语义主要蕴涵于单词当中，而对于中文，由于其语义主要蕴涵于短语中，需要对文档进行词的切分，经过多方面的研究，现在已经有了一些比较好的分词方法，尽管它们还有不够完善的地方，但在大多数情况下还是工作的比较好的。采用分词方法进行汉语 feature 的提取，我们就可以将下列方法直接应用于中文信息的检索。

1）RSV 法（Retrival Status Value）

这是一个比较经典的方法，它实际上是将文档表示成向量形式，向量的分量与 feature j_i 在文档中的出现频度以及该 feature 的代表性等有关（包含 feature j_i 的文档越多，其代表性越差，反之，则越高）。对查询串 query 也作同样的处理，然后计算 query 与文档 d_j 的向量表示的夹角余弦值，即 RSV（q，d_j）= $r_{\cos}(\vec{q},\vec{d}_j)$，最后按该值从大到小的顺序向用户返回查询结果。

根据这种方法：

- 查询串中的 feature 在文档 d_j 中出现的越多，RSV（q，d_j）越高。

- 查询串中某 feature 在文档d_j中出现的频度越高，RSV（q，d_j）越高。
- 查询串中的 feature j_i 在各文档中出现的越普遍，则它对 RSV（q，d_j）的贡献越小。

可以看出，这种方法不仅计算上简单，而且定性分析的结果也比较合理，所以得到了较为普遍的使用。

2）聚类法

采用该方法时，系统先将各文档表示成RSV法中所述的feature向量表示，并进行规范化，然后根据这些向量表示的点在空间上的距离进行聚类。当进行查询时，查询串也被表示成 feature 向量形式并规范化，然后在空间上选择离它最近的文档类作为搜索的基本范围，并在此范围内使用具体的检索方法，最终得到查询结果。尽管这种方法有一定的问题，如加入新文档后的算法稳定性、文档类数目的预确定等。但它的使用能大大缩小文档的搜索空间，减少每次查询的计算工作量，所以随着它的继续完善，这种方法还是很有前途的。

3）相关矩阵法

定义矩阵 $X_0 = [\vec{d}_1\ \vec{d}_2 \ldots \vec{d}_j \ldots \vec{d}_m]$，$\vec{d}_j = (a_{1j}, a_{2j}, \ldots, a_{ij}, \ldots, a_{nj})$，$a_{ij} = ff(j_i, d_j)$。$ff(j_i, d_j)$表示 feature j_i 在文档d_j中的频度。研究发现，存在这样的一些j_i，它们在各文档中的出现都比较频繁，由于它们在各文档中的普遍出现，使得人们不得不怀疑它们是某种表达或句法结构上的需要，而在描述文档内容方面却无多大的意义，这样的j_i称为 feature 噪声。

当 feature 噪声很大时，会对相关度的计算产生很大的干扰，使查询结果精确度下降，所以对它的消除显得非常重要，前述的 RSV 法中的$idf(j_i)$因子的引入是一种行之有效的方法，这里我们要介绍的是另一种方法，它对X_0进行分解：$X_0 \rightarrow T_0 S_0 D_0^T$，其中$S_0$是对角阵。在将$S_0$中较小的特征值去除后，得到$S_1$，而$T_0$、$D_0$也将相应行或列去除得到$T_1$、$D_1$，然后将它们合并：$T_1 S_1 D_1^T \rightarrow X_1$，所得$X_1$即为相关矩阵（详细过程请见参考文献【2】）。其实$X_0$到$X_1$的转变过程正是 feature 噪声的消除过程，那些在X_0中表现很强烈的噪声j_k在X_1中会很微弱，由S_0到S_1是消除噪声的关键所在。

得到X_1后就可用 RSV 等方法继续进行处理了。

4）双字匹配法

这是一种针对汉语的方法。它的依据是汉语中绝大部分词汇是双字的，因此在该方法中不需要进行分词，而只要将查询串或文档切分成一个个连续的双字，然后再通过匹配和词频统计计算相关度。这种方法会产生很多无意义的词，但影响并不大，举个例子：当用户查询“古代历史”时，假设有两篇文章 A 和 B，A 中含有“古代历史”，B 中仅有“古代”和“历史”这两个分开的串。经过切词后，查询串会被切成“古代”、“代历”和“历史”，显然“代历”是个无意义的词，但它在查询串和 A 中都出现了，所以按词频统计所得的查询串与 A 的相关度会大于它与 B 的相关度，因此 A 会被优先选中。这种方法实现简单、无需使用复杂的分词算法，而且其查询效果与基于分词的检索方法相比，并无明显差异，因此具有很强的实用性。东南大学与 ETH 合作的中文搜索引擎“网达”是采用此方法的典型。

在使用索引检索方法时，可以考虑相关性查询以丰富查询结果，比如当用户查询“布”时，可以不仅返回有关明指“布”的文档，还可以返回与“丝绸”、“涤纶”等有关的文档。这种方法需要搜索引擎具有更强的字典和分类方面的功能等。

可以看出，基于索引的检索方法与全文检索方法相比，有很大的优势，由于它首先对查询串和文档进行 feature 的提取，然后再进行匹配，使得该方法在一定程度上具有语义查询的特点，较之全文检索的简单机械的匹配方法效果要好很多，而且对较长的查询串也能有不错的查询结果。

3．3 其它值得考虑的问题

1）交叉语言查询

有时用户希望在多语种间查询信息，或者他希望某语种的信息，却不能用该语言表达他的查询要求，这种情况下，进行交叉语言的查询就显得很有必要。一般说来，交叉查询的方法有对查询串的机器翻译、计算不同语种文档的相关性等等。这是目前研究的一个热点。

2）交互式搜索

这是一个策略上而非技术上的问题。在设计一个搜索引擎时，应该将它与用户看成一个整体。许多搜索引擎都向用户提供了一个反馈的机会，因为有时用户的信息需求是很难用有限长度的查询串表达清楚的，所以查询串不一定很准确，而且搜索引擎的搜索结果有时也不一定很另人满意，这样，让用户在查询结果里向系统反馈他最感兴趣的文档就很有必要。系统可以根据这些反馈信息，调整查询串并进行新的查询，将一批类似的文档提供给用户。这样的过程重复下去，会在很大程度上弥补技术的不足，提高用户的满意程度。

3）集成服务的提供

未来的发展趋势是，一个搜索引擎不仅仅是基于内容的网上信息的查询，它还应包括特殊查询（如人名、地名）、多信息源混合查询（如 www、ftp、gopher）、多媒体查询等多种服务。这样，才能吸引更多的用户，提高竞争力。

4．体系结构方面的考虑

从体系结构上看，现有的搜索引擎一般都是集中式的。通常它们之间并没有什么协作，而是各自独立地搜集和处理信息，这造成了大量重复工作和网络带宽的严重浪费，同时也给各 WEB 站点带来了沉重的负担。所以这种体系结构难以适应网络规模的日益扩大。下面的分布式体系结构能有效解决这些问题。

分布式体系结构的基本思想是：将全网划分成若干自治域，这种自治域的划分根据可以是地理空间，也可以是 IP 地址或信息内容。在每个自治域内分别设立一个 gatherer 和 broker，gatherer 负责本自治域信息的搜集，broker 负责的工作比较复杂，如向用户提供查询接口，与 gatherer 进行交互，信息的处理入库等，更重要的，它必须与其它 broker 进行交互，例如，当用户向某 broker 提交查询串时，该 broker 除了可以在本地查询外，还可以根据需要将查询串提交给其它的 broker，从而获得其它自治域的查询结果。当然，为使各 broker 协同一致的工作，它们之间还要进行一些控制信息的传递。采取这种分布式体系结构，各 gatherer 只负责本自治域内的信息搜集，各 WEB 站点也只向本自治域内的 gatherer 提供信息，从而减轻了网络负载和各站点的负载，同时各 broker 间的互助协作关系使得服务的提供更为灵活和完善，而且该体系结构使得系统的维护和扩充变得更为方便。

进一步地，我们还可以在自治域内部的各 WEB 站点上设立服务器端搜集代理，其功能是负责本地信息的摘录，当 gatherer 搜集信息时，它按最近更新的原则将所有可提供的信息摘录与其它一些信息如文档 URL 等组成具有一定结构的数据流一次发送给 gatherer。这样，一方面无需传送文档全文，另一方面各站点也无需为每个文档建立传送连接，从而进一步地减轻了网络负载和各站点的负载。

5．结束语

因特网的飞速发展，向人类提供了一个巨大而丰富的信息源泉，但由于其运行机制的特殊性和以 WWW 为代表的网上信息服务的极度分布性，使得网上的信息导航功能变得越来越重要，于是以搜索引擎为核心的各种信息导航手段应运而生。这种以应用为驱动所产生的服务需求是跨越语言边界的，但由于起步较晚，许多中文搜索引擎与国外的西文搜索引擎相比，还很不成熟。在本文中我们就中文搜索引擎的信息检索和体系结构等问题，从技术角度进行一些分析和探讨，以期能够更好地促进这项技术的发展，充分满足广大汉语用户的需求。

参考文献

【1】刘挺，吴岩，王开铸，“串频统计和词形匹配相结合的汉语自动分词系统”，《中文信息学报》，98 年第 1 期

【2】Christos Faloutsos，Douglas Oard “A survey of Information retrieval and filtering method”

【3】C.MicBowman ，Peter B.Danzig，etc. “Harvest: A scalable，Customizable Discovery and Access System”

【4】Pedro Fall · Gon，Silvio lemos Meira，etc.“A distributed Mobile Code-Based Architecture for Information Indexing，Searching，and Retrieval on the World Wide Web”

Researches on Chinese Information Search Engine within INTERNET

Shen Yujiang Ding Wei
(Computer Department, Southeast University, Nanjing, 210096)

【Abstract】The paper studies Chinese Search Engine techniques in network environment from two different points of view which influence the performance seriously. The first one is about information retrieval skills. Several ways for Chinese information retrieval are introduced, analyzed and compared with each other. The last one is about the architecture of the search engine. Distributed solutions are arisen and some other relevant ideals, such as search agent, are involved also.

【Key Words】Search Engine, Chinese Information Retrieval, Computer Network, Distributed Architecture.