基于 WWW 的多媒体信息搜索系统

杨惠　（湖北大学数学与计算机科学学院　武汉 430062）

摘要：本文介绍了一个多媒体信息搜索系统 TIMSE(text - image meta - search engine)用于实现 WWW 上大型、分布式联机信息系统的检索。它可以搜索 Web 上的文本及图像信息，并将用户反馈结合到一个性能排序机制中。

关键词：WWW　信息检索　多媒体信息　性能数据库　用户反馈

1.引言

随着 WWW(World Wide Web)上信息爆炸式地增长，对于用户来说要寻找到他们感兴趣的信息变得越来越困难。因此，各类帮助用户寻找他们感兴趣的信息的检索工具变得非常有用和流行，于是在 WWW 上出现了各种各样的搜索引擎，例如 AltaVIsta, Excite Infosee, Lycos 和 Hotbot。然而大多数搜索引擎都有一些局限性。

·首先，这些搜索引擎都是文本性的。只要给出几个关键词，它们就可以检索出含有这些关键词的 Web 文件来。尽管许多 Web 文件都含有图像，但对于用户来说，指定一个图像并检索出含有相似图像的 Web 文件可不是一件容易的事。

·其二，许多搜索引擎依赖 robot 去发现联机发布的信息，但是由于 WWW 的内容太多而不可能在单个服务器上都检索到。即使这样做，这单个服务器也会因为要为世界各地的查询提供服务而过载，而且网络的时间延迟和信息通信量也显著增加。

·最后，由于大多数用户对于他们检索的领域或索引数据库的关键词不太了解，因此在他们的查询语句中提供的信息很少，相关性的评估效率不高。

检索服务的局限性导致了元搜索引擎 meta - search engine(如 MetaCrawler 和 SavvySearch)的出现。Meta - search engine 通过将查询要求传给多个搜索引擎如 AltaVista 或 Excite 来搜索 WWW。目前 meta - search engine 的主要优点是它能综合多个搜索引擎的搜索结果，返回许多以前不能发现的 Web 文件，并提供统一的用户界面来访问这些搜索引擎 。

2. TIMSE 系统基本的体系结构

多媒体信息搜索引擎 TIMSE(Text - image meta search engine)是一个搜索 Web 上文本和图像信息资源的元搜索引擎。为了获得联机的文本和图像信息资源，它能自动地将用户连接到多个文本或图像搜索引擎。TIMSE 系统的总体结构见图 1 所示。

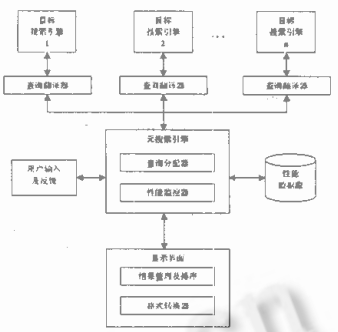

图 1　TIMSE 系统基本的体系结构

图中该系统的三个主要部件是标准的元搜索引擎的部件：查询分配器(Query dispatcher)，查询翻译器(Query translater)及显示界面组件(Display interface component)。

·一旦接受到查询要求，查询分配器通过咨询 TIMSE 的性能数据库(Performance database)，选择目标搜索引擎进行查询。性能数据库包含有系统所支持的搜索选项的以前查询的成功与失败的性能分数。

·接着，查询翻译器将用户的查询要求翻译成适合所选的目标搜索引擎界面的脚本。

·最后，显示界面组件对目标搜索引擎返回结果进行整理与聚类，重新排序，并将最终结果显示给用户。

·用户对显示结果做出反馈，可进一步求精查询，指导系统的搜索过程，以便更准确地检索到所需的信息。

TMISE 根据用户的反馈来评估每个搜索选项的返回结果质量。这一信息可用来修改性能数据库里相应的属性项。

TIMSE 设计的一个基本原则就是保证返回结果质量的同时，尽可能有效地利用 Web 的资源。这一原则在系统实现的许多方面体现。例如，在类似于过去查询的情况下，只查询那些能提供较好查询结果的搜索引擎。

TIMSE 系统可以用 C 语言来实现。TIMSE 用套接(Socket)程序与单个目标搜索引擎进行通信，但这一方式对用户来说是透明的。HTTP 命令传送给远程的搜索引擎，提出查询要求，以类似于 Netscape 和 Mosaic 浏览器的方式下载它们的结果。TIMSE 系统支持文本和图像搜索引擎。其中

文本搜索引擎有：Altavista、Lycos、Yahoo、Infoseek。

图像搜索引擎有：VisualSeek、WebSeek、QBIC、Virage。

这些搜索引擎都有它们自己特有的功能及局限性。

Altavista、Infoseek 和 Yahoo 支持由多个词组成的短语搜索，只有这些词连续同时出现才算是正确的结果。Altavista 和 Infoseek 可以以某个特定栏目上的内容来搜索，如页面上的标题、URL 地址、电子邮件等。Altavista、Infoseek 和 Lyeos 的搜索结果按搜索词出现的频繁与否给定的分数的大小排序，而 Yahoo 则不行。Altavista 可以用逻辑运算符(and、or、not)来表示每个词或短语间的关系。

VisualSeek，Virage 和 QBIC 接受例图像，根据图像特征如颜色和纹理将实例图像与库里图像相匹配。VisualSeek 和 QBIC 支持自定义搜索，允许用户交互式地绘出视图轮廓或指定外部图像，而 Virage 允许用户权衡搜索时每个图像特征的重要性。WebSeek 是一个半自动化的图像和视频搜索引擎。它使用颜色直方图以支持基于文本的搜索。

TIMSE 用独特的方式将文本及图像两个检索系统结合到一起。它允许用户搜索出含有一个或多个关键字及类似于用户指定图像的 Web 文件。TIMSE 将 Harvest 信息发现和访问系统用于检索文本索引与搜索，而 COIR(content Oriented Image Retrival)面向内容图像检索库用于图像的搜索。

3. TIMSE 的索引(indexing)

TIMSE 的索引机制如图 2 所示。在 TIMSE 的 Web 服务器里，Harvest 的搜索数据库用于收集目标搜索引擎返回的 HTML 文件。当 Harvest 完成 HTML 文件的收集后，搜索数据库就可以提供 HTML 文件所含或所指的图像信息。用被检索出的图像的数目。他也可以改变 color－texture 比率，此比率在确定颜色和纹理相似性时很重要。

如果文本和图像查询使用了 and 逻辑符号，那么只有被文本和图像搜索引擎都检索到的 Web 文本才能显示给用户。检索出的文本显示用关键词的匹配行的数目来排序，而图像则用总的匹配率来排序。

最后，TIMSE 系统对搜索结果实现可视化。可视化是由可产生虚拟现实模拟语言 VRML(Virtual Reality Modeling Language)码的程序来实现。

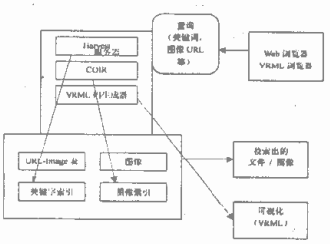

图 2　TIMSE 的查询机制

(Harvest 和 COIR 用于检索符合特定查询的 Web 文件，VRML 可视化查询结果)

4. TMISE 的性能数据库及用户查询

当 TIMSE 接受到查询要求时，查询分配器就选择目标搜索引擎及查询的搜索选项。所谓搜索选项即是在特定的搜索引擎上的查询方法。首先，查询分配器根据提交给 TMISE 的查询类型来做出选择。其中，目标搜索引擎的选择是基于性能数据库里的分数。这些分数记录着过去每次查询时每个搜索选项的执行情况。在 TIMSE 中，基于关键字的文本查询是由一个或多个关键字来确定。一旦接受到一个查询要求时，TIMSE 就搜索性能数据库，检索出查询的性能分数。查询分配器将选出符合用户关键字、图像特征及语义类别并带有最高分数的搜索选项。

(下转第 14 页)

(上接第 16 页)

如果提出了一个以前从未有的新查询,这样数据库里就不可能有它的性能分数,则最简单的方法就是随机地选出搜索选项进行查询。然而,TIMSE 却是通过将新查询与过去类似的查询相比较来推荐搜索选项。性能数据库里的查询根据查询内容被聚合成(Clustering)十几类。当用户提出一个新查询时,系统就下载查询内容(关键字及图像),将它匹配到相应的聚类结构以获得最相似的聚类列表,这样查询分配器就可根据最相似的聚类的性能分数推荐出合适的搜索引擎来。同时,这个新查询将被加到性能数据库中以便将来查询时使用。

参考文献

[1] 翁智泓,信息搜索技巧,南开大学出版社;1998.

[2] 曹明,Web 信息搜索方法,计算机与通讯;1997.4

[3] 曹莉华等,基于 WWW 的多媒体信息检索,微型电脑应用;1998.4

(来稿时间:1999 年 3 月)