目　录

关于比较分析 Web 3.0 搜索引擎的综述报告

摘　要.....1

关键词.....1

I　对搜索引擎的介绍.....1

II　对 Web 3.0 的介绍.....3

III　标注方法.....4

IV　几乎不被使用的 WEB 3.0 搜索引擎.....4

译者：钟明媛

时间：2015 年 4 月 25 日

备注：英文版论文原文置于译文前边，页码仅针对译文。

感谢翻译过程中老师与好心朋友的指点。

因译者能力有限，若有不足，请谅解，并愿意接受指点。

关于比较分析 Web 3.0 搜索引擎的综述报告

R.Aravindhan，助理教授
计算机科学与工程系
斯里兰卡 eshwar 工程学院
印度，泰米尔纳德邦，哥印拜陀
contact2aravind@gmail.com

R.Shanmugalakshmi 博士，副教授
计算机科学与工程系
信息技术政府科技学院
印度，泰米尔纳德邦，哥印拜陀
drshanmuhslskdhmi@yahoo.co.in

摘 要 通过无数数据库在互联网上所累积的信息量是极其庞大的。而从互联网上对这些信息的搜索正是由人们所熟知的专业工具——搜索引擎来完成的。搜索引擎基于用户所输入的关键词来搜索所需信息，它可被视为一款简单的的软件程序。通常这些搜索与检索是基于对关键词的语法分析（Web 2.0），但是如果其是基于内容分析，那么反馈给用户的将是一个更有意义的结果。语义万维网（Web 3.0）则将是当前网络的一种扩展，它的关键词将被赋予更明确的含义。除此之外，语义万维网的另一优势是它那友好的用户界面。最近，大量以不同设计原则为基础和为用户或应用程序提供不同支持程度的语义网搜索引擎已经在不断出现。然而，尽管这些搜索引擎拥有广泛有效的直觉力，为用户提供了好处，但是用户仍是不情愿转向使用这些高端的搜索引擎。在本论文中，我们通过对语义搜索引擎的比较分析调查，以此呈现语义搜索引擎那前景广阔的特性以及找出用户为何难以接受使用这些高端搜索引擎的背后原因。

关键词 Web 3.0，语义搜索引擎，语义网，信息检索，网络爬虫

I 对搜索引擎的介绍

[1]互联网搜索引擎独特且重要，网络用户通过专门的网站搜索遍布各网站的重要信息。这个将每个搜索引擎联系起来的宏伟计划与设计将主要开展以下三项工作：

i）根据关键词搜索互联网并获取信息。

ii）通过保留关键词的索引来定位关键词的起点。

iii）允许用户利用索引查询词或短语。[1]

[978-1-4799-2352-6/14/$31.00 ©2014 IEEE]

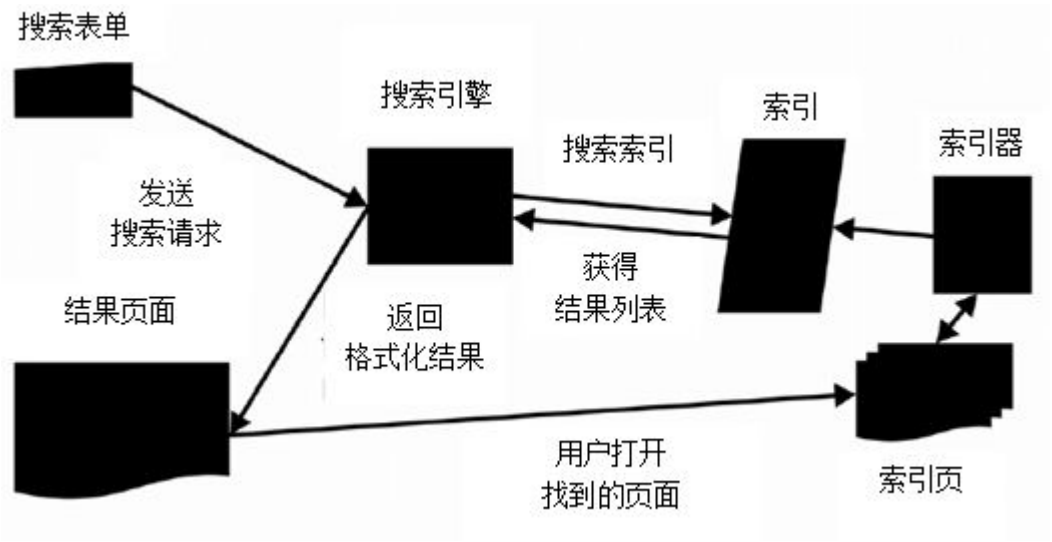

图 1 搜索引擎的工作流程

搜索引擎的工作流程如*图-1* 所示。它由以下步骤组成：

a.建立索引；

b.接受搜索查询（一套搜索词和指令）；

c.搜寻索引文件，进行匹配；

d.汇集匹配网页条目，并根据其与关键词的关联程度进行排列；

e.格式化搜索结果；

f.以 HTML 的版式将页面数据反馈到用户的浏览器。

搜索引擎也被视为一组安排有序的计算机程序：（a）如*图-1* 中被称作“网络爬虫”或“机器人”的搜索引擎蜘蛛，像蜘蛛在蜘蛛网上爬行一样，在所有网页中从一个链接自动爬到另一个链接。（b）程序把蜘蛛爬行过的网页建立起一个目录索引。（c）当程序接收到一个搜索请求，会与目录中的关键词进行比较，然后将搜索结果反馈到用户的搜索引擎。

[2]通常网页内容中缺乏一个合适的结构，该结构与呈现在网页上的信息相关，因此，传统的搜索引擎其自身有如下局限性：

1. 预期搜索信息与检索信息的互连程度不佳。

2. 处理数量庞大的网络用户和网页内容时，无法保证各级可信度。

3. 提供的信息缺乏理解力（可解释为难以理解）。

4. 信息自动传递能力不足。

II 对 Web 3.0 的介绍

Web 1.0 主要致力于搜索，然而 Web 2.0 主要参与到社会活动中并与其进行交互，Web 3.0 则是基于服务。[3]Web 3.0（语义网络）一直向将搜索信息与检索信息联系起来的方向演化。它能够查询互联网上的所有信息，让电脑理解网上的信息，从而替我们执行日益复杂的任务。

Web 3.0 被理解为“物联网”。每个独立事物都具有：（a）一个名字和网络上的地址。（b）一个已经创建的人们可读的页面。（c）那些机器可读的数据，这些数据是用来描述事物所想要的和可提供的。（d）Web 3.0 也被称作“服务网络”。网络服务设定程序，去读有关事物的机器可读数据，以及利用数据链接将机器的需求与相关事物联系起来。

Web 2.0 与 Web 3.0 的比较

表 1

	Web 2.0	**Web 3.0**
权限	可读/可写	可移植 Web
用户	群体/社区	个体
数据	共享数据	合并动态内容
技术	AJEX	RDF
举例	Google,Wikipedia	Dbpedia,igoogle

语义网的一个重要目标是想要让被搜索的内容的实际含义更加明确，因此力求在异构信息环境中使有效的信息获取成为可能。语义搜索在实现这个目标上发挥着重要的作用，它利用明确的信息语义的可用性，努力确保能够反馈给用户与其请求相关的结果。

当代的网络搜索引擎不会精确地解释人们打算搜索的内容。它只不过是通过用户输入的关键词在网页上进行搜索和检索。除非网页返回的内容是用户实际上期望的，否则这搜索引擎最主要的缺点就是不能确保搜索结果的准确度。比如，一个用户想要知道水果苹果的价格，于是输入关键词“苹果的价格”，而返回的结果中包含了苹果手机的价格。

通过 Web 3.0 浏览器，你会发现一些关于你搜索的关键词和内容。该浏览器除了会返回一些相关的结果，同时还会提供给你与关键词相关的其他内容。比如，当你在 Web 3.0 搜索引擎输入“低于 3000 美元的热带旅游景区”，随即会出现海量与关键词有关的内容，例如若干娱乐活动的介绍或者许多附近酒店的推荐。每一个在 Web 3.0 搜索引擎提交的查询请

求，都会让搜索引擎在整个网站内容中进行大量的搜索，从而返回给用户最有用最相关的信息。

III 标注方法

在日益增长的诸多问题中，研究人员们在 SW 领域正遭遇一个叫标注的事物，标注是 SW 搜索引擎中的一个强制性要求。通过让合适的元数据作为前置条件的办法，设法让当前网页能够采取 SW 搜索去标注它。一般而言，给结构化数据生成元数据会相对简单些。

可以根据以下几个方面给标注分类：

A.元数据类型：根据[25]元数据可以被分成两种类型，如结构化和语义化。在结构化元数据中，没有上下文信息的表达（如语言和格式）。然而在语义化元数据中，不仅有上下文信息的表达，而且通常以 RDF 三元组的形式存储。

B.生成方法：在生成过程中，生成方法十分简单，无需考虑网页的所有主题，只需要利用一个网页的结构化信息或者本体即可生成元数据。这方法的主要优点：基于上下文信息生成元数据。

C.生成的来源：通常元数据生成的来源是网页本身，但是有时候利用其它补充的资源也是有益的。比如[1]和[5]从可以利用的网络中使用信息，为自己的网页累积更多的信息。例如在搜索一部电影时，会提取到一些像导演、流派等附加信息。

IV 几乎不被使用的 WEB 3.0 搜索引擎

Hakia

Hakia 搜索引擎：搜索包括词汇的含义和由三种科学知识开发方法所组成的语句。

I. OntoSem(感觉库)

II.QDEX(查询索引法)

III.语义排序算法

OntoSem 是一类语言数据库。每个用户所说的词汇被描述或被追踪成不同“感觉”，OntoSem 正与这些想法有关。

QDEX 伴随着查询检测和提取倒排索引而发展着，该索引是用来在更多引擎之上存储网页内容用的。所以，QDEX 正在逐渐取代 Hakia 搜索引擎。最后，它不断获取与网页内容可能相关的问题，并采用语义排序算法对基于语句分析的内容进行排序。

图 2 hakia 搜索引擎

Hakia 搜索引擎有以下几个特征[6]：

- 致力于高质量与精确的结果。
- 可以获取与数据相关的信息。
- 在用户不允许的情况下，不会将数据保存在用户的系统中。
- 在最短时间内得到目标信息，以此使用户感到高兴。
- 设法让搜索者在首次尝试查询时获得更多精确且相关的结果。
- 基于查询，Hakia 返回分类的结果。
- Hakia 理解一词多义和同义，并提供检索的相关信息。

Factbites:

[7]Factbites 是另外一款搜索引擎，它根据搜索查询的主题，检索出有意义的结果。它不是通过匹配关键词来罗列出诸多网站，而是剖析关键词的含义，检索出更精确更明确的结果。在主题和用户的关键词两者之中，该款搜索引擎倾向于通过主题来匹配内容。它尝试为用户提供更真实更有意义的内容。

[8]Factbites 通过进行简洁的搜索使得网络搜索更有效，比如给用户一个有意义的、相关的响应。它不仅仅只是进行关键词匹配，或是关注有多少关键词已经匹配，同时还要保证网页能够正确地返回用户所需的信息。在撤掉网站列表之前，它会分析已经被检索过的网页的含义，过滤垃圾信息网站。

图 3 factbites 搜索引擎

Factbites 搜索引擎有以下几个特征：

- 提供附加的真实资料。
- 可以在搜索结果中过滤掉垃圾信息网站。
- 根据主题进行搜索胜于根据关键词。