网络爬虫技术学术发展分析报告

Analysis Report on Academic Development of Web Reptile Technology

方建勇[1]（余姚，浙江 315400）

摘要：通过超星发现系统，我们大致了解到网络爬虫技术所涉及的相关领域，这些领域的研究课题属于哪些学科，哪些机构发表的学术成果较多，集中在哪些刊物发表等信息，还有指出了哪些网络爬虫技术的相关学术成果被引用较多，为我们研究网络爬虫技术这个课题做了比较好的指引。

关键词：网络爬虫技术 搜索引擎 URL 模式 主题爬虫 分析报告

Abstract： Through the superstar discovery system, we have a general understanding of the relevant areas involved in web crawler technology, which disciplines of these research topics, which institutions have published more academic results, focused on which publications and other information, and which networks Reptile technology related academic results are cited more for our study of web crawler technology to do a better guide to this topic.

Key words： Web crawler technology; Search engine; URL pattern; Theme crawler; Analysis report

随着 Internet 的迅速发展,网络资源越来越丰富,人们如何从网络上抽取信息也变得至关重要,尤其是占网络资源 80%的 Deep Web 信息检索更是人们应该倍加关注的难点问题[2]。为提高主题爬虫的性能，依据站点信息组织的特点和 URL 的特征，提出一种基于 URL 模式集的主题爬虫。爬虫分两阶段，在实验爬虫阶段，采集站点样本数据，采用基于 URL 前缀树的模式构建算法构建 URL 模式，形成模式关系图，并利用 HITS 算法分析该模式关系图，计算出各模式的重要度；在聚焦爬虫阶段，无须预先下载页面，即可利用生成的 URL 模式判断页面是否主题相关和能否指导爬虫深入抓取，并根据 URL 模式的重要度预测待抓取链接优先级。实验表明，该爬虫相比现有的主题爬虫能快速引导爬虫抓取主题相关页面，

---

[1] 方建勇，男，1978 年-，中国工业与应用数学学会会员，中国物流学会会员，中国计算机学会会员，浙江大学历史系硕士研究生学历，浙江大学数学与应用数学专业本科毕业，理学学士学位。

[2] 曾伟辉，李淼（中国科学院合肥智能机械研究所；中国科学院合肥智能机械研究所；中国科学技术大学自动化系），深层网络爬虫研究综述[J]，《计算机系统应用》 2008 第 17 卷 第 5 期 P122-122 。

保证爬虫的查准率和查全率，有效提高爬虫抓取效率。[3]

本文谨对网络爬虫技术的学术情况作一个基于大数据的分析，希望能对研究能有所帮助。爬虫，也是爬行动物的一种，本文检索到的此类学术成果，考虑到数量不大，且网络爬虫也是由此引申而来，故予以了保留。

一、网络爬虫技术学术发展趋势

超星发现系统收录的网络爬虫技术历年发表的学术成果，见表 1，总量为 7,824 条记录，包括中文 6953 条和外文 871 条 。其中包括期刊 (3520)、报纸 (265)、学位论文(2198)、会议论文 (163)、标准 (3)、专利 (1011)、音视频 (302)、科技成果 (88)、年鉴 (20)、法律法规 (7)、案例 (3)、信息资讯 (129)、特色库 (4) 。

表 1 网络爬虫技术各类型学术发展趋势

爬虫-各类型学术发展趋势									
序号	年份	图书(数量)	期刊(数量)	学位论文(数量)	会议论文(数量)	专利(数量)	标准(数量)	报纸(数量)	科技成果(数量)
1	1921	0	1	0	0	0	0	0	0
2	1922	0	0	0	0	0	0	0	0
3	1923	0	1	0	0	0	0	0	0
4	1924	0	0	0	0	0	0	0	0
5	1925	0	0	0	0	0	0	0	0
6	1926	0	0	0	0	0	0	0	0
7	1927	0	1	0	0	0	0	0	0
8	1928	0	2	0	0	0	0	0	0
9	1929	0	0	0	0	0	0	0	0
10	1930	1	1	0	0	0	0	0	0
11	1931	0	0	0	0	0	0	0	0
12	1932	1	0	0	0	0	0	0	0
13	1933	0	0	0	0	0	0	0	0
14	1934	0	0	0	0	0	0	0	0
15	1935	0	1	0	0	0	0	0	0
16	1936	0	1	0	0	0	0	0	0

[3] 胡萍瑞，李石君（武汉大学 计算机学院），基于 URL 模式集的主题爬虫[J]，《计算机系统应用研究》，2018 第 3 期 。

17	1937	0	0	0	0	0	0	0	0
18	1938	0	0	0	0	0	0	0	0
19	1939	0	0	0	0	0	0	0	0
20	1940	0	1	0	0	0	0	0	0
21	1941	0	0	0	0	0	0	0	0
22	1942	0	1	0	0	0	0	0	0
23	1943	1	1	0	0	0	0	0	0
24	1944	0	0	0	0	0	0	0	0
25	1945	0	0	0	0	0	0	0	0
26	1946	0	0	0	0	0	0	0	0
27	1947	0	0	0	0	0	0	0	0
28	1948	0	0	0	0	0	0	0	0
29	1949	0	0	0	0	0	0	0	0
30	1950	0	1	0	0	0	0	0	0
31	1951	0	0	0	0	0	0	0	0
32	1952	0	1	0	0	0	0	0	0
33	1953	0	0	0	0	0	0	0	0
34	1954	0	2	0	0	0	0	0	0
35	1955	0	1	0	0	0	0	0	0
36	1956	1	0	0	0	0	0	0	0
37	1957	0	0	0	0	0	0	0	0
38	1958	0	2	0	0	0	0	0	0
39	1959	1	2	0	0	0	0	0	0
40	1960	0	0	0	0	0	0	0	0
41	1961	0	1	0	0	0	0	0	0
42	1962	0	4	0	0	0	0	0	0
43	1963	0	1	0	0	0	0	0	0
44	1964	1	23	0	0	0	0	0	0
45	1965	0	1	0	0	0	0	0	0
46	1966	0	2	0	0	0	0	0	0
47	1967	0	0	0	0	0	0	0	0
48	1968	0	11	0	0	0	0	2	0
49	1969	0	0	0	0	0	0	0	0
50	1970	0	2	0	0	0	0	0	0
51	1971	0	6	0	0	0	0	0	0
52	1972	0	21	0	0	0	0	0	0
53	1973	0	1	0	0	0	0	0	0
54	1974	0	5	0	0	0	0	0	0
55	1975	0	28	0	0	0	0	0	0

56	1976	0	5	0	0	0	0	0	0
57	1977	0	29	0	0	0	0	0	0
58	1978	0	2	0	0	0	0	0	0
59	1979	0	25	0	0	0	0	0	0
60	1980	0	7	0	0	0	0	0	0
61	1981	0	30	0	0	1	0	0	0
62	1982	1	8	0	0	0	0	0	0
63	1983	0	51	0	0	0	0	0	0
64	1984	0	11	0	0	0	0	0	0
65	1985	0	42	0	0	0	0	0	0
66	1986	0	4	0	0	0	0	0	0
67	1987	1	41	0	0	1	0	0	0
68	1988	0	19	0	0	0	0	0	0
69	1989	0	38	0	1	0	0	0	0
70	1990	0	7	0	0	1	0	0	0
71	1991	0	43	0	0	4	0	0	0
72	1992	1	7	0	0	1	0	0	0
73	1993	0	30	0	0	1	0	0	1
74	1994	0	15	0	0	2	0	0	0
75	1995	0	50	0	0	0	0	0	0
76	1996	0	12	0	1	1	0	0	0
77	1997	0	35	0	0	2	0	0	0
78	1998	0	17	0	0	0	0	1	0
79	1999	1	32	0	2	3	1	0	0
80	2000	1	36	1	0	0	0	0	0
81	2001	1	22	1	0	1	0	2	0
82	2002	0	50	3	1	1	0	0	1
83	2003	0	27	2	1	4	0	2	1
84	2004	0	32	7	2	5	0	1	0
85	2005	1	56	14	2	3	0	1	1
86	2006	2	78	31	4	18	0	4	0
87	2007	4	127	74	9	20	0	13	0
88	2008	7	154	113	18	28	0	20	3
89	2009	8	143	142	19	21	0	14	2
90	2010	6	252	181	27	38	0	45	3
91	2011	6	232	242	22	60	1	21	5
92	2012	10	254	227	9	66	0	32	7
93	2013	7	242	264	9	92	0	37	1
94	2014	12	289	300	17	134	1	34	2

95	2015	11	315	310	10	199	0	12	2
96	2016	9	305	233	9	250	0	10	0
97	2017	12	293	174	0	295	0	13	0

二、网络爬虫技术学术成果统计[4]

1、关键词

关键词涉及网络爬虫(845)、搜索引擎(522)、垂直搜索引擎(156)、爬行动物(149)、数据挖掘(133)、lucene(122)、微博(121)、网络舆情(108)、垂直搜索(101)、中文分词(96)、文本分类(91)、设计实现(91)、本体(90)、信息抽取(84)、向量空间模型(79)、信息检索(77)、信息采集(74)、hadoop(71)、互联网(64)、系统设计(61)、自然保护区(61)、搜索策略(56)、社交网络(55)、分布式(51)、特征提取(49)、两栖动物(48)、文本挖掘(46)、WEB(45)、数据采集(42)、机器学习(39)、支持向量机(37)、链接分析(37)、电子商务(36)、技术研究(36)、聚类(34)、文本聚类(34)、野生动物(33)、web 挖掘(33)、脊椎动物(33)、DeeP Web(33)、云计算(32)、遗传算法(30)、Pagerank(30)、Mapreduce(30)、生物多样性(29)、多线程(29)、资源调查(29)、全文检索(29)、个性化(28)等。

[4] 数据来源于超星发现系统。

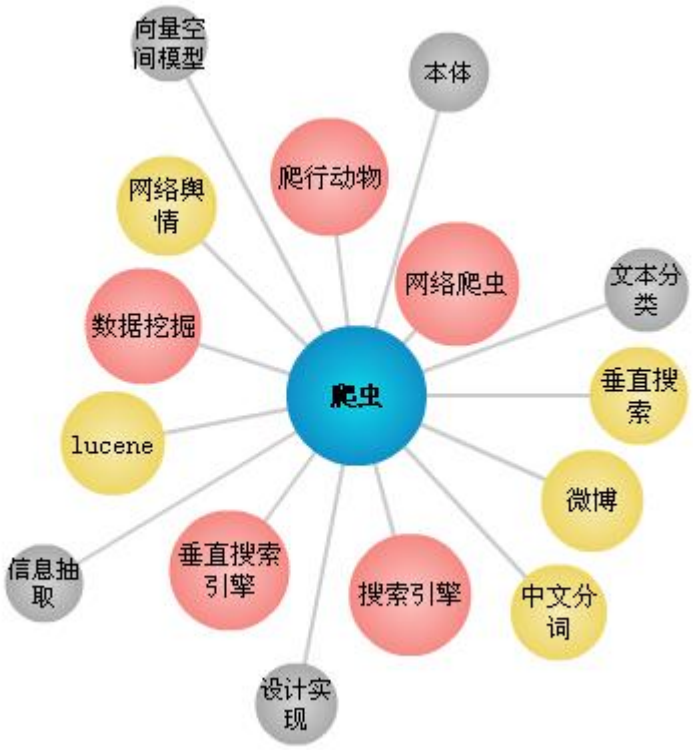

图 1 网络爬虫技术关键词频次泡型图

2、机构、刊种与地区分布

发表机构分布依次为北京邮电大学(209)、中国科学院(132)、电子科技大学(128)、哈尔滨工业大学(127)、华中科技大学(100)、北京航空航天大学(99)、中山大学(75)、武汉大学(69)、北京大学(66)、中国人民大学(58)、北京交通大学(58)、华南理工大学(55)、吉林大学(53)、浙江大学(49)、武汉理工大学(49)、西安电子科技大学(49)、厦门大学(46)、北京理工大学(45)、复旦大学(44)、上海交通大学(43)、北京工业大学(40)、南京邮电大学(40)、四川大学(39)、中国科学院大学(39)、南京大学(38)、苏州大学(36)、华中师范大学(35)、同济大学(34)、东南大学(32)、西南交通大学(32)、大连理工大学(30)、山东大学(30)、华东师范大学(27)、天津大学(26)、南京师范大学(26)、重庆邮电大学(26)、南京理工大学(25)、重庆大学(24)、西北工业大学(24)、国防科学技术大学(24)、湖南大学(23)、清华大学(22)、北京林业大学(22)、东北大学(22)、中国科学技术大学(22)、湖北工业大学(22)、中南林业科技大学(22)、广东工业大学(22)、西安交通大学(22) 等。

发表的刊物分布依次为爬虫両棲類学雑誌(357)、爬虫両棲類学会報(195)、ク

リーパー：爬虫両生類情報誌(149)、电脑知识与技术(49)、爬蟲兩棲類學雜誌(36)、计算机应用与软件(33)、Current herpetology(32)、计算机技术与发展(32)、BULLETIN OF THE HERPETOLOGICAL SOCIETY OF JAPAN(31)、计算机工程(31)、软件导刊(25)、计算机工程与设计(24)、动物学杂志(24)、计算机系统应用(23)、计算机应用(23)、北海道爬虫両棲類研究報告(21)、信息网络安全(20)、计算机应用研究(20)、四川动物(18)、古脊椎动物学报(16)、计算机与现代化(16)、安徽农业科学(16)、现代图书情报技术(15)、计算机科学(14)、软件(14)、电脑编程技巧与维护(14)、计算机光盘软件与应用(13)、计算机工程与科学(13)、现代计算机(13)、科技信息(12)、电脑知识与技术(学术交流)(12)、水生生物学报(11)、计算机工程与应用(11)、水族世界(11)、微计算机信息(11)、林业调查规划(11)、动物学报(英文版)(11)、科学技术创新(10)、计算机与数字工程(10)、模型世界(10)、情报杂志(10)、生态学报(10)、计算机时代(9)、信息与电脑(理论版)(9)、电子世界(8)、生物多样性(7)、网络安全技术与应用(7)、数字技术与应用(7)、古生物学报(7)等。

发表机构所属的地区分布依次是北京市(823)、湖北省(320)、江苏省(317)、广东省(280)、四川省(250)、上海市(236)、黑龙江省(181)、陕西省(164)、浙江省(131)、辽宁省(117)、山东省(116)、湖南省(113)、安徽省(99)、吉林省(89)、福建省(83)、重庆市(71)、天津市(67)、河南省(60)、甘肃省(56)、江西省(53)、云南省(53)、河北省(45)、广西壮族自治区(42)、新疆维吾尔自治区(34)、贵州省(30)、内蒙古自治区(26)、海南省(23)、山西省(19)、西藏自治区(3)、青海省(2)、宁夏回族自治区(2) 等。

3、网络爬虫技术期刊论文

超星发现系统收录的网络爬虫技术学位论文 853 篇，网络爬虫技术高引前 10 篇期刊论文见表。

表 2 网络爬虫技术高引前 10 篇期刊论文列表

作者	学术成果	被引用次数	发表年份	类型	所在机构	刊名
周立柱，	聚焦爬虫技术研究综述	382	2005	期刊	清华大学计算机	计算机应用

林玲					科学与技术系	
刘金红，陆余良	主题网络爬虫研究综述	270	2007	期刊	解放军电子工程学院网络系	计算机应用研究
周德懋，李舟军	高性能网络爬虫：研究综述	151	2009	期刊	北京航空航天大学计算机学院	计算机科学
汪涛，樊孝忠	主题爬虫的设计与实现	141	2004	期刊	北京理工大学计算机科学工程系,中国人民解放军炮兵学院三系	计算机应用
孙立伟，何国辉，吴礼发	网络爬虫技术的研究	120	2010	期刊	解放军理工大学指挥自动化学院	电脑知识与技术
郑冬冬，赵朋朋，崔志明	Deep Web 爬虫研究与设计	90	2005	期刊	苏州大学计算机科学与技术系	清华大学学报(自然科学版)
李勇，韩亮	主题搜索引擎中网络爬虫的搜索策略研究	89	2008	期刊	大连海事大学计算机科学与技术学院	计算机工程与科学
徐远超，刘江华，刘丽珍，关永	基于 Web 的网络爬虫的设计与实现	81	2007	期刊	首都师范大学信息工程学院	微计算机信息杂志
曾伟辉，李淼	深层网络爬虫研究综述	66	2008	期刊	中国科学院合肥智能机械研究所	计算机系统应用
郑冬冬，崔志明	Deep Web 爬虫爬行策略研究	56	2006	期刊	苏州大学智能信息处理及应用研究所	计算机工程与设计

4、网络爬虫技术专项研究学位论文

超星发现系统收录的网络爬虫技术学位论文 403 篇，可以说是代表了当前网络爬虫技术水平最高的群体，网络爬虫技术高引前 30 篇学位论文见表 3。

表 3　网络爬虫技术高引前 30 篇学位论文列表

作者	学术成果	发表年份	类型	所在机构
罗兵	支持 AJAX 的互联网搜索引擎爬虫设计与实现	2007	硕士	浙江大学
刘玮玮	搜索引擎中主题爬虫的研究与实现	2006	硕士	南京理工大学
朱良峰	主题网络爬虫的研究与设计	2008	硕士	南京理工大学
程锦佳	基于 Hadoop 的分布式爬虫及其实现	2010	硕士	北京邮电大学
苏旋	分布式网络爬虫技术的研究与实现	2006	硕士	哈尔滨工业大学
曾伟辉	支持 AJAX 的网络爬虫系统设计与实现	2009	硕士	中国科学技术大学
沈寿忠	基于网络爬虫的 SQL 注入与 XSS 漏洞挖掘	2009	硕士	西安电子科技大学
苏晓珂	基于 Nutch 的主题爬虫研究与实现	2007	硕士	昆明理工大学
陈丛丛	主题爬虫搜索策略研究	2009	硕士	山东大学
袁小节	基于协议驱动与事件驱动的综合聚焦爬虫研究与实现	2009	硕士	国防科学技术大学
杨贞	基于本体的主题爬虫的设计与实现	2008	硕士	合肥工业大学
叶勤勇	基于 URL 规则的聚焦爬虫及其应用	2007	硕士	浙江大学
倪贤贵	聚焦爬虫技术研究	2008	硕士	江南大学
李玉华	面向主题的舆情采集搜索爬虫的设计与实现	2009	硕士	山东大学
蒋科	基于领域概念定制的主题爬虫系统的设计与实现	2007	硕士	西安电子科技大学

郑博文	基于 Hadoop 的分布式网络爬虫技术	2011	硕士	哈尔滨工业大学
么士宇	基于分布式计算的网络爬虫技术的研究	2011	硕士	大连海事大学
龚勇	搜索引擎中网络爬虫的研究	2010	硕士	武汉理工大学
王桂梅	主题网络爬虫关键技术研究	2009	硕士	哈尔滨工业大学
杨溥	搜索引擎中爬虫的若干问题研究	2009	硕士	北京邮电大学
龚秋艳	并行网络爬虫设计与实现	2010	硕士	华东师范大学
张航	主题爬虫的实现及其关键技术研究	2010	硕士	武汉理工大学
贺晟	搜索引擎中主题网络爬虫的研究与设计	2010	硕士	安徽大学
袁浩	主题爬虫搜索 Web 页面策略的研究	2009	硕士	中南大学
杜一平	主题搜索网络爬虫的设计与研究	2009	硕士	中国科学技术大学
刘喜亮	面向主题的网络爬虫设计与实现	2009	硕士	湖南大学
刘洁清	网站聚焦爬虫研究	2006	硕士	江西财经大学
梁萍	搜索引擎中网络爬虫及结果聚类的研究与实现	2011	硕士	中国科学技术大学
林碧霞	基于领域本体的主题爬虫研究及实现	2010	硕士	西南交通大学
夏亮	主题搜索引擎网络爬虫搜索策略的研究与实现	2010	硕士	北京化工大学

三、结语

通过超星发现系统，我们大致了解到网络爬虫技术所涉及的相关领域，这些领域的研究课题属于哪些学科，哪些机构发表的学术成果较多，集中在哪些刊物发表等信息，还有指出了哪些网络爬虫技术的相关学术成果被引用较多，为我们研究网络爬虫技术这个课题做了比较好的指引。

四、参考文献

[1] 超星发现系统[EB/OL].http://www.chaoxing.com/

[2] 曾伟辉，李淼（中国科学院合肥智能机械研究所；中国科学院合肥智能机械研究所；中国科学技术大学自动化系），深层网络爬虫研究综述《计算机系统应用》， 2008 第 17 卷 第 5 期 P122-122 。
[3] 胡萍瑞，李石君（武汉大学 计算机学院），基于 URL 模式集的主题爬虫[J]，《计算机系统应用研究》，2018 第 3 期 。