



(12) 发明专利申请

(10) 申请公布号 CN 103559257 A

(43) 申请公布日 2014. 02. 05

(21) 申请号 201310535301. X

(22) 申请日 2013. 11. 01

(71) 申请人 深圳市宜搜科技发展有限公司
地址 518026 广东省深圳市福田区滨河路
与彩田路交汇处联合广场 A 栋塔楼
A5501-A

(72) 发明人 庞伟

(74) 专利代理机构 深圳市凯达知识产权事务所
44256
代理人 任转英 刘大弯

(51) Int. Cl.
G06F 17/30 (2006. 01)

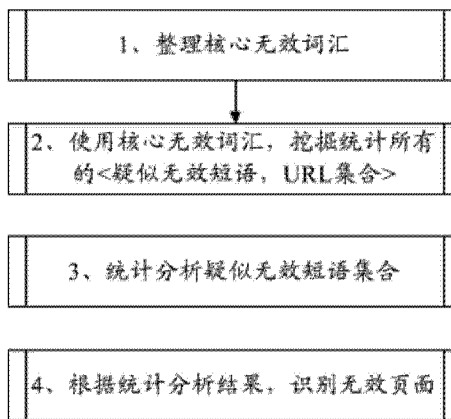
权利要求书1页 说明书7页 附图3页

(54) 发明名称

一种互联网无价值网页的挖掘方法

(57) 摘要

本发明提供一种互联网无价值网页的挖掘方法,包括:步骤1、整理核心无效词汇;步骤2、使用核心无效词汇,识别出所有的疑似无效短语;步骤3、统计分析疑似无效短语集合;步骤4、根据统计分析结果,识别无效页面。通过本发明提供的一种互联网无价值网页的挖掘方法,通过对网页聚类、统计分析的手段,解决了互联网网页中无效页面的高准确率、高召回率、自动化的挖掘问题,从而提高挖掘速度、改善搜索引擎的用户体验。



1. 一种互联网无价值网页的挖掘方法,其特征在于,包括:

步骤 1、整理核心无效词汇;

步骤 2、使用核心无效词汇,识别出所有的疑似无效短语;

步骤 3、统计分析疑似无效短语集合;

步骤 4、根据统计分析结果,识别无效页面。

2. 如权利要求 1 所述的方法,其特征在于,步骤 1 中,具体包括:一个短句是无效短语的必要条件是含有最普遍出现的核心无效词汇。

3. 如权利要求 2 所述的方法,其特征在于,步骤 2 中,具体包括:对于基于单页面级别的无效页面挖掘,在打开的网页,建立 DOM 树,遍历纯文本节点,纯文本长度阈值在 10 字节和 20 字节之间且含有至少一个核心无效词汇,则该纯文本称为疑似无效短语。

4. 如权利要求 2 所述的方法,其特征在于,步骤 2 中,具体包括:对于基于站点级别的无效页面挖掘:

在打开的网页,建立 DOM 树,遍历纯文本节点,纯文本长度阈值在 10 字节和 20 字节之间且含有至少一个核心无效词汇,则该纯文本称为疑似无效短语;

将疑似无效短语和该网页的 URL 构成一个映射对:〈疑似无效短语,URL〉。

5. 如权利要求 3 所述的方法,其特征在于,步骤 3 中,具体包括:对于基于单页面级别的无效页面挖掘,统计分析疑似无效短语集合。

6. 如权利要求 4 所述的方法,其特征在于,步骤 3 中,具体包括:对于基于站点级别的无效页面挖掘:

将〈疑似无效短语,URL〉聚类:以疑似无效短语为键值,将 URL 映射到一个簇内,形成一个含有相同疑似无效短语的 URL 集合聚簇;

统计分析〈疑似无效短语,URL 集合〉的聚簇:URL 集合按照 URL 字典序排序,含有最长公共子串的一串 URL 的数量超过阈值 200 个,并且占聚簇的比重超过阈值 70%,构成一个〈URLPattern,无效短语〉,其中,URLPattern 是最长公共子串;

根据统计分析的聚簇得到一系列无效模式对〈URLPattern,无效短语〉。

7. 如权利要求 6 所述的方法,其特征在于,所述最长公共子串 URLPattern 的长度至少是协议头的长度与站点长度之和,且必须超过阈值 12 字节。

8. 如权利要求 5 所述的方法,其特征在于,步骤 4 中,具体包括:对于基于单页面级别的无效页面挖掘,当疑似无效短语集合的统计分析结果为以下之一时,即可判页面为无效页面:

疑似无效短语集合总数量超过阈值 20,即可判为无效页面;

疑似无效短语集合总数量介于 5 和 20 之间,且不同的疑似无效短语数量超过阈值 5,即可判为无效页面。

9. 如权利要求 6 所述的方法,其特征在于,步骤 4 中,具体包括:对于基于站点级别的无效页面挖掘:

加载无效模式对〈URLPattern,无效短语〉;

一个新网页 URL,如果匹配 URLPattern,且页面内一个独立的 DOM 树纯文本节点含有与 URLPattern 配对的无效短语,这个网页即可判为无效页面。

一种互联网无价值网页的挖掘方法

技术领域

[0001] 本发明涉及互联网搜索领域,特别是涉及一种互联网无价值网页的挖掘方法。

背景技术

[0002] 互联网海量的网页信息,网页的质量参差不齐,搜索引擎要爬取、存储、索引。识别、过滤无价值的网页,可节省搜索引擎的存储,提高索引的数据质量,改善用户的搜索体验。

[0003] 互联网网页,尤其是生活服务、电子商务等时效性较强的网页,容易变成无效页面,这类页面对用户没有价值,影响用户的搜索体验。

[0004] 无效页面,是指网页内容中含有“无效短语”的页面。而无效短语,是指用户的核心检索需求没有满足、没找到、不存在等内容的短语。例如,“该商户点评暂停收录”、“暂时还没有任何榜单”、“暂时还没有任何内容”、“暂时还没有任何相册”、“该商户已关闭”等无效短语。网页中含有此类短语,均视作无效页面,如图 2、图 3、图 4、图 5 所示。

[0005] 传统的无效页面挖掘方法,首先人工整理收集一批无效短语;对每一个爬取的网页,所有的无效短语都参与匹配,匹配上至少一个无效短语即判定为无效页面。这种传统的人工整理方法虽然简单,但存在明显的不足:人工方法准确率低,误判率高(达到 20% 到 30%);浪费人力资源;整理的无效短语数量有限;不能实现自动化挖掘。

[0006] 所以,如何保证高准确率、高召回率的挖掘无效短语、挖掘无效页面,是搜索引擎必须解决的数据质量问题。

发明内容

[0007] 本发明提供一种互联网无价值网页的挖掘方法,通过对网页聚类、统计分析的手段,解决了互联网网页中无效页面的高准确率、高召回率、自动化的挖掘问题,从而提高挖掘速度、改善搜索引擎的用户体验。

[0008] 为了解决以上技术问题,本发明提供一种互联网无价值网页的挖掘方法,包括:

[0009] 步骤 1、整理核心无效词汇;

[0010] 步骤 2、使用核心无效词汇,识别出所有的疑似无效短语;

[0011] 步骤 3、统计分析疑似无效短语集合;

[0012] 步骤 4、根据统计分析结果,识别无效页面。

[0013] 进一步地,步骤 1 中,具体包括:一个短句是无效短语的必要条件是含有最普遍出现的核心无效词汇,

[0014] 进一步地,步骤 2 中,具体包括:对于基于单页面级别的无效页面挖掘,在打开的网页,建立 DOM 树,遍历纯文本节点,纯文本长度阈值在 10 字节和 20 字节之间且含有至少一个核心无效词汇,则该纯文本称为疑似无效短语。

[0015] 进一步地,步骤 2 中,具体包括:对于基于站点级别的无效页面挖掘:在打开的网页,建立 DOM 树,遍历纯文本节点,纯文本长度阈值在 10 字节和 20 字节之间且含有至少一

个核心无效词汇,则该纯文本称为疑似无效短语;将疑似无效短语和该网页的 URL 构成一个映射对:〈疑似无效短语,URL〉。

[0016] 进一步地,步骤 3 中,具体包括:对于基于单页面级别的无效页面挖掘,统计分析疑似无效短语集合。

[0017] 进一步地,步骤 3 中,具体包括:对于基于站点级别的无效页面挖掘:

[0018] 将〈疑似无效短语,URL〉聚类:以疑似无效短语为键值,将 URL 映射到一个簇内,形成一个含有相同疑似无效短语的 URL 集合聚簇;

[0019] 统计分析〈疑似无效短语,URL 集合〉的聚簇:URL 集合按照 URL 字典序排序,含有最长公共子串的一串 URL 的数量超过阈值 200 个,并且占聚簇的比重超过阈值 70%,构成一个〈URLPattern,无效短语〉,其中,URLPattern 是最长公共子串;

[0020] 根据统计分析的聚簇得到一系列无效模式对〈URLPattern,无效短语〉。

[0021] 进一步地,所述最长公共子串 URLPattern 的长度至少是协议头的长度与站点长度之和,且必须超过阈值 12 字节。

[0022] 进一步地,步骤 4 中,具体包括:对于基于单页面级别的无效页面挖掘,当疑似无效短语集合的统计分析结果为以下之一时,即可判页面为无效页面:

[0023] 疑似无效短语集合总数量超过阈值 20,即可判为无效页面;

[0024] 疑似无效短语集合总数量介于 5 和 20 之间,且不同的疑似无效短语数量超过阈值 5,即可判为无效页面。

[0025] 进一步地,步骤 4 中,具体包括:对于基于站点级别的无效页面挖掘:

[0026] 加载无效模式对〈URLPattern,无效短语〉;

[0027] 一个新网页 URL,如果匹配 URLPattern,且页面内一个独立的 DOM 树纯文本节点含有与 URLPattern 配对的无效短语,这个网页即可判为无效页面。

[0028] 与相对传统的无效页面挖掘方法相比,本发明提供了一种互联网无价值网页的挖掘方法,紧扣无效页面本身特征,通过对网页聚类、统计分析的手段,解决了互联网网页中无效页面的高准确率、高召回率、自动化的挖掘问题,从而实现了自动化挖掘,极大化准确率与召回率,提高无价值信息挖掘速度和效果、改善搜索引擎的用户体验。经过测试实验,在基于站点级别的无效页面挖掘中,测试实验的准确率在 98%,召回率在 65% 以上;在基于单页面级别的无效页面挖掘中,准确率超过 95%,召回率超过 60%。

[0029] 此外,本发明提供的站点级别和页面级别两个无效页面挖掘算法均易于程序实现,可周期性执行挖掘应对无效短语时效性特点。

附图说明

[0030] 此处所说明的附图用来提供对本发明的进一步理解,构成本发明的一部分,本发明的示意性实施例及其说明用于解释本发明,并不构成对本发明的不当限定。在附图中:

[0031] 图 1 是本发明提供一种互联网无价值网页的挖掘方法的流程图。

[0032] 图 2 是某网页出现含有“暂时还没有任何内容”无效短语的截图。

[0033] 图 3 是某网页出现含有“该商户点评暂停收录”无效短语的截图。

[0034] 图 4 是某网页出现 28 个无效短语的截图。

[0035] 图 5 是某网页出现 27 个无效短语的截图。

具体实施方式

[0036] 为了使本发明所要解决的技术问题、技术方案及有益效果更加清楚、明白,以下结合附图和实施例,对本发明进行进一步详细说明。应当理解,此处所描述的具体实施例仅用以解释本发明,并不用于限定本发明。

[0037] 一般来说,挖掘互联网无价值网页,核心难点在于挖掘无效短语。无效短语的一般特征,主要包括有:

[0038] 特征 A、无效短语具有地域性。一个无效短语并不适用于所有的网页,限定在其高频出现的特定范围内,某一个站点下、某一层 URL 路径下,构成一个无效模式对 <URLPattern, 无效短语>, 无效页面识别的准确率与召回率会极大化。

[0039] URLPattern, 是指某一个站点下的某一层 URL 目录。例如 :http://www.dianping.com/shop 就是一个 URLPattern。

[0040] 地域性的表现之一就是 :无效短语都是单独出现的,在网页的 DOM 树节点中独占一个纯文本节点,且均出现在网页中央主体位置。

[0041] 特征 B、无效短语具有时效性。这表现在 :无效短语易变,一段时间过去后会发生变化;不同行业的网页含有不同的无效短语,例如,生活服务类的网页与电子商务类的网页,其无效短语不同。

[0042] 在本发明中,无效短语的挖掘,分为站点级别和页面级别,将两者分开的目的是最大化准确率、召回率。

[0043] 如图 1 所示,本发明提供一种互联网无价值网页的挖掘方法,包括:

[0044] 步骤 1、整理核心无效词汇;具体包括:

[0045] 一个短句是无效短语的必要条件是含有最普遍出现的核心无效词汇,比如:“暂停”、“关闭”、“过期”、“无效”、“封禁”、“暂无”、“找不到”等。

[0046] 步骤 2、使用核心无效词汇,识别出所有的疑似无效短语;具体包括:

[0047] 在基于单页面级别的无效页面挖掘中,在打开的网页,建立 DOM 树,遍历纯文本节点,纯文本长度阈值在 10 字节和 20 字节之间并且含有至少一个核心无效词汇,这个纯文本称为疑似无效短语。

[0048] 在基于站点级别的无效页面挖掘中,在打开的网页,建立 DOM 树,遍历纯文本节点,纯文本长度阈值在 10 字节和 20 字节之间并且含有至少一个核心无效词汇,这个纯文本称为疑似无效短语;将疑似无效短语和该网页的 URL 构成一个映射对 :<疑似无效短语, URL>。

[0049] 步骤 3、统计分析疑似无效短语集合,得到统计分析结果;具体包括:

[0050] 基于单页面级别的无效页面挖掘中,统计分析疑似无效短语集合,得到统计分析结果。

[0051] 基于站点级别的无效页面挖掘中,具体包括:

[0052] 将 <疑似无效短语, URL> 聚类 :以疑似无效短语为键值 key,将 URL 映射到一个簇内,形成一个含有相同疑似无效短语的 URL 集合聚簇。

[0053] 统计分析 <疑似无效短语, URL 集合> 的聚簇 :URL 集合按照 URL 字典序排序。很显然,一个站点下的 URL 会顺序出现。同时含有最长公共子串的一串 URL 的数量超过阈值

200 个,并且占聚簇的比重超过阈值 70%,就构成一个 <URLPattern,无效短语>。最长公共子串就是 URLPattern。最长公共子串长度必须超过阈值 12 字节,最长公共子串长度至少是协议头的长度与站点长度之和 [sizeof(http://)+sizeof(site)]。

[0054] 根据统计分析的聚簇得到一系列无效模式对 <URLPattern,无效短语>。

[0055] 步骤 4、根据统计分析结果,识别无效页面;具体包括:

[0056] 基于单页面级别的无效页面挖掘中,当疑似无效短语集合的统计分析结果为以下之一时,即可判页面为无效页面:疑似无效短语集合总数量超过阈值 20,即可判为无效页面;若疑似无效短语集合总数量介于 5 和 20 之间,且不同的疑似无效短语数量超过阈值 5,即可判为无效页面。

[0057] 基于站点级别的无效页面挖掘中,具体包括:加载无效模式对 <URLPattern,无效短语>。一个新网页 URL,如果匹配 URLPattern,并且页面内一个独立的 DOM 树纯文本节点含有与 URLPattern 配对的无效短语,这个网页即可判为无效页面。

[0058] 在本发明中,无效短语挖掘,分为站点级别和页面级别,将两者分开的目的是最大化准确率、召回率。下面将结合实施例来详细说明本发明在站点级别和页面级别的实施方式,借此对本发明如何应用技术手段来解决技术问题,并达成技术效果的实现过程能充分理解并据以实施。

[0059] 一:基于站点级别的无效页面挖掘

[0060] 本发明实施例提供一种基于站点级别的互联网无效页面的挖掘方法,包括:

[0061] 步骤 A1:整理核心无效词汇

[0062] 一个短句是无效短语的必要条件是含有最普遍出现的核心无效词汇,例如:“暂停”、“关闭”、“过期”、“无效”、“封禁”、“暂无”、“找不到”等。本发明中整理出以下核心无效词汇。例如:暂停、关闭、错误、不存在、故障、禁止、不正确、抱歉、返回、无法、暂无、删除、重试、不真实、请输入、对不起、找不到、非法、下架、提示、是否、跳转、有误、无权、过期、稍候、失败、暂时、无效、失效、重新、再继续、不支持、请谅解、仅支持、还没有、没有权限、没有记录、暂时没有、没有找到、没有符合、没有发表、维护中、请不要、尚未有、您还不是、正在审核、锁定、封禁、暂不等等。

[0063] 步骤 A2、使用核心无效词汇,识别出所有的疑似无效短语

[0064] 在打开的网页,建立 DOM 树,遍历纯文本节点,纯文本长度阈值在 10 字节和 20 字节之间并且含有至少一个核心无效词汇,这个纯文本称为疑似无效短语。

[0065] 疑似无效短语和该网页的 URL 构成一个映射对:<疑似无效短语,URL>。

[0066] 步骤 A3、统计分析疑似无效短语集合,挖掘得到一系列无效模式对 <URLPattern,无效短语>;具体包括:

[0067] 步骤 A3.1:<疑似无效短语,URL> 聚类

[0068] 以疑似无效短语为键值 key,将 URL 映射到一个簇内,形成一个含有相同疑似无效短语的 URL 集合聚簇。

[0069] 步骤 A3.2:统计分析 <疑似无效短语,URL 集合> 的聚簇

[0070] URL 集合按照 URL 字典序排序。很显然,一个站点下的 URL 会顺序出现。同时含有最长公共子串的一串 URL 的数量超过阈值 200 个,并且占聚簇的比重超过阈值 70%,就构成一个 <URLPattern,无效短语>。最长公共子串就是 URLPattern。最长公共子

串长度必须超过阈值 12 字节,最长公共子串长度至少是协议头的长度与站点长度之和 [sizeof(http://)+sizeof(site)]。

[0071] 步骤 A3.3:根据 A3.2 的统计分析结果,得到一系列的无效模式对 <URLPattern, 无效短语 >。

[0072] 一个 URLPattern,可能对应着若干个无效短语,一个具体的网页实际上只含有其中的一个无效短语。例如图 2 和图 3 所示:

[0073] <http://www.dianping.com/member 暂时还没有任何榜单 | 暂时还没有任何点评 | 暂时还没有任何粉丝 | 暂时还没有任何关注 | 暂时还没有任何内容 | 暂时还没有任何签到 | 暂时还没有任何荣誉 | 暂时还没有任何相册 >

[0074] <http://www.dianping.com/shop 该商户点评暂停收录 | 评分:该商户暂无星级 >

[0075] 对于这种情况,可适当调整阈值的大小,平衡准确率与召回率。提高阈值,倾向于准确率;降低阈值,倾向于召回率。

[0076] 步骤 A4:使用无效模式对 <URLPattern, 无效短语 >,识别无效页面。

[0077] 加载无效模式对 <URLPattern, 无效短语 >。一个新网页 URL,如果匹配 URLPattern,并且页面内一个独立的 DOM 树纯文本节点含有与 URLPattern 配对的无效短语,这个网页即可判为无效页面。测试实验的准确率在 98%,召回率在 65% 以上。

[0078] 基于站点级别的无效页面挖掘的结果,得到一系列无效模式对 <URLPattern, 无效短语 >,符合特征 A,某一个站点下、某一层 URL 路径下高频出现的无效短语,这一类页面是无效页面的准确率较高。

[0079] 二:基于单页面级别的无效页面挖掘

[0080] 单个页面含有的比较多且稀有的无效短语,且其它网页不含有这些无效短语,那么上述统计方法失效。基于站点级别的挖掘方法不能召回此类无效页面。

[0081] 例如图 4 所示。图 4 中含有的 28 个无效短语:暂无索县内容!| 暂无索县内容!| 暂无索县最新信息 | 暂无索县竞价信息 | 暂无索县推荐信息 | 暂无索县热门信息 | 暂无索县内容!| 暂无索县内容!| 暂无索县内容!| 暂无索县内容!| 暂无索县内容!| 暂无索县用户留言!| 暂无索县信息!| 暂无索县信息!| 暂无索县信息!| 暂无索县信息!| 暂无索县信息!| 暂无索县信息!| 暂无索县信息!| 暂无索县信息!| 暂无索县信息!| 暂无索县信息!| 暂无索县内容!| 暂无索县图片链接!| 暂无索县文字链接!。

[0082] 图 4 和图 5 这两个页面含有的无效短语稀有,在其它页面没有,所以统计无效短语频数失效。

[0083] 本发明实施例提供一种基于单页面级别的互联网无效页面的挖掘方法,包括:

[0084] 步骤 B1:整理核心无效词汇,整理方法与站点级别的相同。

[0085] 步骤 B2:使用核心无效词汇,识别出所有的疑似无效短语。

[0086] 在打开的网页,建立 DOM 树,遍历纯文本节点,纯文本节点长度在 10 字节和 20 字节之间且含有至少一个核心无效词汇,这是疑似无效短语,保存起来。

[0087] 步骤 B3:统计分析疑似无效短语集合。

[0088] 步骤 B4:根据统计分析结果,识别无效页面。

[0089] 当疑似无效短语集合的统计分析结果为以下之一时,即可判页面为无效页面:疑

似无效短语集合总数量超过阈值 20,即可判为无效页面;若疑似无效短语集合总数量介于 5 和 20 之间,且不同的疑似无效短语数量超过阈值 5,即可判为无效页面。准确率超过 95%,召回率超过 60%。

[0090] 下面将结合实施例来详细说明本发明的实施方式,借此对本发明如何应用技术手段来解决技术问题,并达成技术效果的实现过程能充分理解并据以实施。

[0091] 实施例 1:

[0092] 本发明实施例 1 提供一种基于站点级别的互联网无效页面的挖掘方法,包括:

[0093] 1、整理一批核心无效词汇

[0094] 暂停、关闭、错误、不存在、故障、禁止、不正确、抱歉、返回、无法、暂无、删除、重试、不真实、请输入、对不起、找不到、非法、下架、提示、是否、跳转、有误、无权、过期、稍候、失败、暂时、无效、失效、重新、再继续、不支持、请谅解、仅支持、还没有、没有权限、没有记录、暂时没有、没有找到、没有符合、没有发表、维护中、请不要、尚未有、您还不是、正在审核、锁定、封禁、暂不等等。

[0095] 2、构造核心无效词汇的 tries 树词典

[0096] 核心无效词汇数量少,用 tries 树匹配速度快,内存空间占用低。将 1 中的核心无效词汇转化成二进制的 tries 词典。程序加载 tries 词典的速度快。

[0097] 3、使用核心无效词汇,识别出所有的疑似无效短语

[0098] 输入的新网页 html 文档,构建 DOM 树,只访问纯文本节点长度阈值在 10 字节和 20 字节之间的独立叶节点,不包括锚文纯文本。

[0099] 将纯文本内容,与 tries 树匹配,搜索是否含有核心无效词汇。若含有,保存<疑似无效短语, URL> 对,以疑似无效短语为键值 key,聚类到一个簇内,核心数据结构可使用 `std::map<std::string[疑似无效短语], std::set<std::string>[URL 集合]>`。

[0100] 4、统计分析<疑似无效短语, URL> 集合

[0101] 在核心数据结构 `std::map<std::string, std::set<std::string>>` 中,每一个疑似无效短语,统计含有相同的最长公共子串的数量,以及这部分 URL 占 URL 集合 `[std::set<std::string> 的大小]` 的比重,若数量、比重均超过给定阈值 200, 70%,疑似无效短语升级为真实的无效短语,生成一个无效模式对<最长公共子串,无效短语>。最长公共子串的长度至少超过 `sizeof(http://)+sizeof(site)`。

[0102] 遍历完核心数据结构后,便得到所有的无效模式对<最长公共子串,无效短语>,即<URLPattern,无效短语>。

[0103] 5、一系列无效模式对,作为模式匹配新网页,检查是否是无效页面

[0104] 匹配方法:首先检查新网页的 URL 是否匹配某一个 URLPattern,若匹配上某一个,再看网页内容是否含有与 URLPattern 配对的无效短语,匹配方法同 3. 方法一致,建立 DOM 树,查看符合阈值条件的纯文本节点。若匹配上,即可判为无效页面。

[0105] 实施例 2:

[0106] 本发明实施例 2 提供一种基于单页面级别的互联网无效页面的挖掘方法,包括:

[0107] 1、整理一批核心无效词汇,方法与基于站点级别的挖掘方法的步骤 1 相同。

[0108] 2、构造核心无效词汇的 tries 树词典,方法与基于站点级别的挖掘方法的步骤 2 相同。

[0109] 3、使用核心无效词汇，识别出所有的疑似无效短语

[0110] 在打开的网页，构建 DOM 树，仅查看符合阈值条件的纯文本节点，找到所有的疑似无效短语，并保存，构成一个疑似无效短语集合。

[0111] 4、分析疑似无效短语集合，并根据分析结果，识别无效页面。

[0112] 本发明提供了一种互联网无价值网页的挖掘方法，紧扣无效页面本身特征，通过对网页聚类、统计分析的手段，解决了互联网网页中无效页面的高准确率、高召回率、自动化的挖掘问题，从而实现了自动化挖掘，极大化准确率与召回率，提高无价值信息挖掘速度和效果、改善搜索引擎的用户体验。经过测试实验，在基于站点级别的无效页面挖掘中，测试实验的准确率在 98%，召回率在 65% 以上；在基于单页面级别的无效页面挖掘中，准确率超过 95%，召回率超过 60%。此外，本发明提供的站点级别和页面级别的两个无效页面挖掘算法均易于程序实现，可周期性执行挖掘应对无效短语时效性特点。

[0113] 上述说明示出并描述了本发明的一个优选实施例，但如前所述，应当理解本发明并非局限于本文所披露的形式，不应看作是对其他实施例的排除，而可用于各种其他组合、修改和环境，并能够在本文所述发明构想范围内，通过上述教导或相关领域的技术或知识进行改动。而本领域人员所进行的改动和变化不脱离本发明的精神和范围，则都应在本发明所附权利要求的保护范围内。

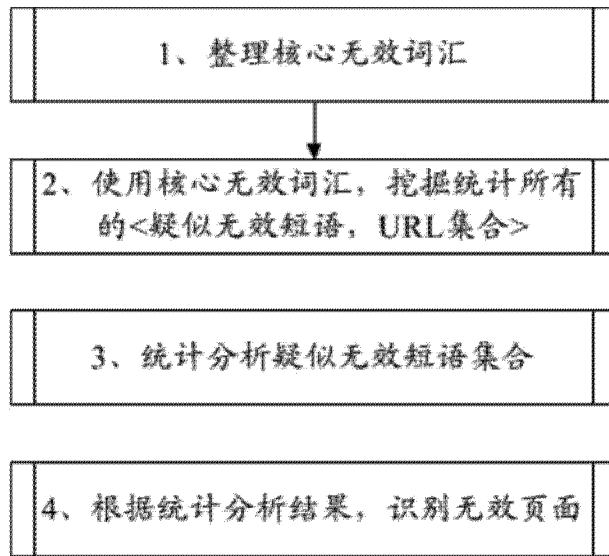


图 1



图 2



图 3



图 4



图 5