

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2017 年 11 月 23 日 (23.11.2017)



(10) 国际公布号
WO 2017/198022 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2017/080784

(22) 国际申请日: 2017 年 4 月 17 日 (17.04.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610321306.6 2016 年 5 月 16 日 (16.05.2016) CN

(71) 申请人: 武汉斗鱼网络科技有限公司
(**WUHAN DOUYU NETWORK TECHNOLOGY CO., LTD.**) [CN/CN]; 中国湖北省武汉市东湖开发区软件园东路 1 号软件产业 4.1 期 B1 栋 11 楼覃波, Hubei 430000 (CN)。

(72) 发明人: 白凡 (**BAI, Fan**); 中国湖北省武汉市东湖开发区软件园东路 1 号软件产业 4.1 期 B1 栋 11 楼覃波, Hubei 430000 (CN)。

(74) 代理人: 武汉智权专利代理事务所 (特殊普通合伙) (**WUHAN ZHI QUAN PATENT AGENCY**); 中国湖北省武汉市东湖新技术开发区珞瑜路 727 号星光无限 4 栋 21 层 2103 室张凯, Hubei 430000 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,

(54) **Title:** LUCENCE CUSTOM THESAURUS-BASED FULL TEXT RETRIEVAL MATCHING METHOD AND SYSTEM

(54) 发明名称: 基于 Lucence 自定义词库的全文检索匹配方法及系统

建立 Lucence 自定义词库: 获取用户输入的搜索词, 检测是否搜索到结果, 搜索不到, 就将搜索词去掉特殊字符, 存入自定义词库; 搜索得到, 就对搜索词进行分词, 对分词后的若干词组继续搜索, 搜索不到, 就去掉特殊字符, 存入自定义词库; 搜索得到, 就记录搜索时间、搜索词、搜索反馈信息

S1

动态分配字段权重: 根据字段的搜索量、搜索反馈信息及自定义权重线性叠加, 按照字段权重动态分配公式定时计算字段权重值, 通过权重设置接口动态赋值给字段

S2

图 1

S1 ESTABLISH A LUCENCE CUSTOM THESAURUS: OBTAIN A SEARCH WORD INPUTTED BY A USER, AND DETECT WHETHER A RESULT IS FOUND; IF A RESULT IS NOT FOUND, REMOVE A SPECIAL CHARACTER FROM THE SEARCH WORD, AND STORE THE WORD IN A CUSTOM THESAURUS; IF A RESULT IS FOUND, SEGMENT THE SEARCH WORD, AND THEN SEARCH THE SEVERAL SEGMENTED WORD GROUPS; REMOVE A SPECIAL CHARACTER FROM A WORD GROUP FOR WHICH A RESULT CANNOT BE FOUND, AND STORE THE WORD GROUP IN THE CUSTOM THESAURUS; IF A RESULT IS FOUND, RECORD A SEARCH TIME, SEARCH WORDS AND SEARCH FEEDBACK INFORMATION

S2 DYNAMIC ALLOCATION OF FIELD WEIGHTS: ACCORDING TO A FIELD SEARCH VOLUME, SEARCH FEEDBACK INFORMATION AND CUSTOM WEIGHTED LINEAR SUPERPOSITION, CALCULATE A FIELD WEIGHT VALUE BY USING A FIELD WEIGHT DYNAMIC ALLOCATION FORMULA IN A TIMED MANNER, AND DYNAMICALLY ASSIGN THE VALUE TO THE FIELD BY MEANS OF A WEIGHT SETTING INTERFACE

(57) **Abstract:** The present invention relates to the field of big data search. A Lucene custom thesaurus-based full text retrieval matching method and system. The method comprises the following steps: obtaining a search word inputted by a user in a real-time manner in a Lucene search environment, detecting whether a result is found; if a result is not found, removing a special character from the search word, and storing the word in a Lucene custom thesaurus; if a result is found, segmenting the search word, continuing to search the several segmented word groups, and detecting whether a result is found; then if a result is not found, removing a special character from the word group for which a result cannot be found, and storing the word group in the Lucene custom thesaurus; if a result is found, recording a search time, segmented search words and search feedback information, and ultimately establishing a Lucene custom thesaurus supporting Lucene full text retrieval. The method is based on the search words inputted by a user, so that a dedicated Lucene custom thesaurus can be established quickly and effectively.

CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 包括国际检索报告(条约第21条(3))。

(57) 摘要：一种基于Lucene自定义词库的全文检索匹配方法及系统，涉及大数据搜索领域。该方法包括以下步骤：**在Lucene搜索环境中，实时获取用户输入的搜索词，检测是否搜索到结果，搜索不到，则将搜索词去掉特殊字符，存入Lucene自定义词库；搜索到，则对搜索词进行分词，对分词后的若干词组继续进行搜索，检测是否搜索到结果，搜索不到，则将搜索不到结果的分词后的词组去掉特殊字符，存入Lucene自定义词库；搜索到，则记录搜索的时间、分词后的搜索词、搜索反馈信息，最终建立支持Lucene全文检索的Lucene自定义词库。该方法根据用户输入的搜索词，能够快速有效地建立自己专用的Lucene自定义词库。**

基于 Lucence 自定义词库的全文检索匹配方法及系统

技术领域

本发明涉及大数据搜索领域，具体是涉及一种基于 Lucence 自定义词库的全文检索匹配方法及系统。

背景技术

Apache Lucence 是一个开放源代码的全文检索引擎工具包，但它不是一个完整的全文检索引擎，而是一个全文检索引擎的架构，提供了完整的查询引擎和索引引擎，部分文本分析引擎。

为了便于读者理解，下面先对相关术语进行简单说明：

Apache Lucence：Apache 旗下一款开源全文检索项目；

全文检索：不同于传统的模糊匹配，先对搜索词按一定规则分词后，再将分词与源数据进行匹配，根据出现分词次数，分词邻近距离，权重等数据进行打分，获得检索结果；

分词：全文检索索引，例如：我是中国人，其分词可为：我、是、中国、人、中国人等；

公共词库：存放公共分词规则的词库，例如，常用的：你好、中国等；

自定义词库：根据自己需求，用于存放自己需要的分词规则的字典词库；

搜索反馈：搜索效果反馈，即用户输入搜索词进入搜索页面后，是否有点击页面链接亦或在多次翻页之后点击链接；

搜索量：某一搜索词某时段全站搜索量；

字段：需要搜索的字段，例如：游戏名、主播名、房间名等。

在 Apache Lucence 全文检索中，需要对源数据进行分词索引。若没有针对特定词组进行分词处理，则该词组无法被检索到。例如，针对游戏直播领域的搜索，“英雄联盟”、“Dota2”、“炉石传说”等在公共词库中基本不会出现的词汇，将很难被检索到。因此，如何获取用户最需要检索的词汇，并生成自定义词库是在全文检索领域一个重要的难点。

发明内容

本发明的目的是为了克服上述背景技术的不足，提供一种基于 Lucence 自定义词库的全文检索匹配方法及系统，根据用户输入的搜索词，能够快速有效地建立自己专用的 Lucence 自定义词库。

本发明提供一种基于 Lucence 自定义词库的全文检索匹配方法，包括以下步骤：

在基于 Lucence 全文检索引擎的搜索环境中，实时获取用户输入的搜索词，检测是否搜索到结果，如果搜索不到结果，则将搜索不到结果的搜索词去掉特殊字符，存入 Lucence 自定义词库；如果搜索到结果，则对搜索到结果的搜索词进行分词处理，得到分词后的若干词组；对分词后的若干词组继续进行搜索，检测是否搜索到结果，如果搜索不到结果，则将搜索不到结果的分词后的词组去掉特殊字符，存入 Lucence 自定义词库；如果搜索到结果，则记录搜索的时间、分词后的搜索词、搜索反馈信息，最终建立支持 Lucence 全文检索的 Lucence 自定义词库。

在上述技术方案的基础上，所述建立支持 Lucence 全文检索的 Lucence 自定义词库之后，还包括以下步骤：在建立支持 Lucence 全文检索的 Lucence 自定义词库的基础上，根据字段的搜索量、搜索反

馈信息及自定义权重变量线性叠加，按照字段权重动态分配公式定时计算字段权重值，再通过 Lucence 全文检索引擎的权重设置接口，将计算得到的字段权重值动态赋值给字段。

在上述技术方案的基础上，所述字段权重动态分配公式为：

$$boost = (\alpha * n + \beta * m + \delta * \ln(t) + r) * \rho,$$

其中，boost 为某字段权重值，n 为某时段该字段被检索量，m 为某时段该字段被检索后具有完整搜索反馈的总量，t 为某时段该字段被检索后不完整搜索反馈的总量，r 为自定义权重变量， α 为检索量的系数因子， β 为完整搜索反馈的系数因子， δ 为不完整搜索反馈的系数因子， ρ 为全局协调系数因子。

在上述技术方案的基础上，所述自定义权重变量为主播名、主播房间名或房间类型。

在上述技术方案的基础上，当系统出现转型或用户搜索喜好发生变化之后，所述自定义权重变量随之改变。

本发明还提供一种基于 Lucence 自定义词库的全文检索匹配系统，该系统包括 Lucence 自定义词库建立单元，所述 Lucence 自定义词库建立单元用于建立支持 Lucence 全文检索的 Lucence 自定义词库：在基于 Lucence 全文检索引擎的搜索环境中，实时获取用户输入的搜索词，检测是否搜索到结果，如果搜索不到结果，则将搜索不到结果的搜索词去掉特殊字符，存入 Lucence 自定义词库；如果搜索到结果，则对搜索到结果的搜索词进行分词处理，得到分词后的若干词组；对分词后的若干词组继续进行搜索，检测是否搜索到结果，如果搜索不到结果，则将搜索不到结果的分词后的词组去掉特殊字符，存入 Lucence 自定义词库；如果搜索到结果，则记录搜索的时间、分词后的搜索词、搜索反馈信息。

在上述技术方案的基础上，该系统还包括字段权重动态分配单元，所述字段权重动态分配单元用于动态分配字段权重：在 Lucence 自定义词库的基础上，根据字段的搜索量、搜索反馈信息及自定义权重变量线性叠加，按照字段权重动态分配公式定时计算字段权重值，再通过 Lucence 全文检索引擎的权重设置接口，将计算得到的字段权重值动态赋值给字段。

在上述技术方案的基础上，所述字段权重动态分配公式为：

$$boost = (\alpha * n + \beta * m + \delta * \ln(t) + r) * \rho,$$

其中，boost 为某字段权重值，n 为某时段该字段被检索量，m 为某时段该字段被检索后具有完整搜索反馈的总量，t 为某时段该字段被检索后不完整搜索反馈的总量，r 为自定义权重变量，例如：主播名、主播房间名、房间类型； α 为检索量的系数因子， β 为完整搜索反馈的系数因子， δ 为不完整搜索反馈的系数因子， ρ 为全局协调系数因子。

在上述技术方案的基础上，所述自定义权重变量为主播名、主播房间名或房间类型。

在上述技术方案的基础上，当系统出现转型或用户搜索喜好发生变化之后，所述自定义权重变量随之改变。

与现有技术相比，本发明的优点如下：

(1) 本发明在基于 Lucence 全文检索引擎的搜索环境中，建立用于进行 Lucence 全文检索的 Lucence 自定义词库，实时获取用户输入的搜索词，检测是否搜索到结果，如果搜索不到结果，则将搜索不到结果的搜索词去掉特殊字符，存入 Lucence 自定义词库；如果搜索到结果，则对搜索到结果的搜索词进行分词处理，得到分词后的若干词组；对分词后的若干词组继续进行检索，检测是否搜索到结果，如

果搜索不到结果，则将搜索不到结果的分词后的词组去掉特殊字符，存入 Lucence 自定义词库；如果搜索到结果，则记录搜索的时间、分词后的搜索词、搜索反馈信息。本发明根据用户输入的搜索词，能够快速有效地建立自己专用的 Lucence 自定义词库，形成符合当前搜索环境的 Lucence 自定义词库，用于 Lucence 全文检索，可达到更佳搜索效果。例如，对于游戏直播，可能用户更期望搜索关于“YYF”、“55 开”、“安德罗妮”等信息，采用常规词库可能无法满足这类需求。采用本发明实施例的方法，在首次搜索时，可能也不会得到最优结果，但是，随着 Lucence 自定义词库不断迭代更新，搜索结果会随着用户的搜索量上涨逐渐优化。

(2) 本发明在 Lucence 自定义词库的基础上，动态分配字段权重：根据字段的搜索量、搜索反馈信息及自定义权重变量线性叠加，按照字段权重动态分配公式定时计算字段权重值，再通过 Lucence 全文检索引擎的权重设置接口 (setboost)，将计算得到的字段权重值动态赋值给字段，能够稳定有效地动态分配各字段权重。当系统出现转型或用户搜索喜好发生变化之后，自定义权重变量随之改变。例如，搜索系统有以下几个字段：主播名、主播房间名、房间类型。系统一开始需要搜索对主播名的搜索更侧重，则只需要增大自定义权重，即字段权重动态分配公式中的自定义权重变量。

附图说明

图 1 是本发明实施例中基于 Lucence 自定义词库的全文检索匹配方法的流程图。

具体实施方式

下面结合附图及具体实施例对本发明作进一步的详细描述。

参见图 1 所示, 本发明实施例提供一种基于 Lucence 自定义词库的全文检索匹配方法, 包括以下步骤:

S1、建立支持 Lucence 全文检索的 Lucence 自定义词库: 在基于 Lucence 全文检索引擎的搜索环境中, 实时获取用户输入的搜索词, 检测是否搜索到结果, 如果搜索不到结果, 则将搜索不到结果的搜索词去掉特殊字符, 存入 Lucence 自定义词库; 如果搜索到结果, 则对搜索到结果的搜索词进行分词处理, 得到分词后的若干词组; 对分词后的若干词组继续进行搜索, 检测是否搜索到结果, 如果搜索不到结果, 则将搜索不到结果的分词后的词组去掉特殊字符, 存入 Lucence 自定义词库; 如果搜索到结果, 则记录搜索的时间、分词后的搜索词、搜索反馈信息, 最终建立支持 Lucence 全文检索的 Lucence 自定义词库;

S2、动态分配字段权重: 在建立支持 Lucence 全文检索的 Lucence 自定义词库的基础上, 根据字段的搜索量、搜索反馈信息及自定义权重变量线性叠加, 按照字段权重动态分配公式定时计算字段权重值, 再通过 Lucence 全文检索引擎的权重设置接口 (setboost), 将计算得到的字段权重值动态赋值给字段;

字段权重动态分配公式为:

$$boost = (\alpha * n + \beta * m + \delta * \ln(t) + r) * \rho,$$

其中, boost 为某字段权重值, n 为某时段该字段被检索量, m 为某时段该字段被检索后具有完整搜索反馈的总量, t 为某时段该字段被检索后不完整搜索反馈的总量, r 为自定义权重变量, 例如: 主播名、主播房间名、房间类型; α 为检索量的系数因子, β 为完整搜索反馈的系数因子, δ 为不完整搜索反馈的系数因子, ρ 为全局协调系数因子。

自定义权重变量可以为主播名、主播房间名或房间类型，当系统出现转型或用户搜索喜好发生变化之后，自定义权重变量随之改变。

本发明实施例还提供一种基于 Lucence 自定义词库的全文检索匹配系统，该系统包括 Lucence 自定义词库建立单元和字段权重动态分配单元，其中：

Lucence 自定义词库建立单元用于建立支持 Lucence 全文检索的 Lucence 自定义词库：在基于 Lucence 全文检索引擎的搜索环境中，实时获取用户输入的搜索词，检测是否搜索到结果，如果搜索不到结果，则将搜索不到结果的搜索词去掉特殊字符，存入 Lucence 自定义词库；如果搜索到结果，则对搜索到结果的搜索词进行分词处理，得到分词后的若干词组；对分词后的若干词组继续进行搜索，检测是否搜索到结果，如果搜索不到结果，则将搜索不到结果的分词后的词组去掉特殊字符，存入 Lucence 自定义词库；如果搜索到结果，则记录搜索的时间、分词后的搜索词、搜索反馈信息；

字段权重动态分配单元用于动态分配字段权重：在建立支持 Lucence 全文检索的 Lucence 自定义词库的基础上，根据字段的搜索量、搜索反馈信息及自定义权重变量线性叠加，按照字段权重动态分配公式定时计算字段权重值，再通过 Lucence 全文检索引擎的权重设置接口（setboost），将计算得到的字段权重值动态赋值给字段；

字段权重动态分配公式为：

$$boost = (\alpha * n + \beta * m + \delta * \ln(t) + r) * \rho,$$

其中，boost 为某字段权重值，n 为某时段该字段被检索量，m 为某时段该字段被检索后具有完整搜索反馈的总量，t 为某时段该字段被检索后不完整搜索反馈的总量，r 为自定义权重变量，例如：主播名、主播房间名、房间类型； α 为检索量的系数因子， β 为完整

搜索反馈的系数因子， δ 为不完整搜索反馈的系数因子， ρ 为全局协调系数因子。

自定义权重变量可以为主播名、主播房间名或房间类型，当系统出现转型或用户搜索喜好发生变化之后，自定义权重变量随之改变。

本发明实施例根据用户输入条件，能够快速有效地建立自己专用的 Lucence 自定义词库，并且形成符合当前搜索环境的 Lucence 自定义词库，用于 Lucence 全文检索，可达到更佳搜索效果。

例如，对于游戏直播，可能用户更期望搜索关于“YYF”、“55开”、“安德罗妮”等信息，采用常规词库可能无法满足这类需求。采用本发明实施例的方法，在首次搜索时，可能也不会得到最优结果，但是，随着 Lucence 自定义词库不断迭代更新，搜索结果会随着用户的搜索量上涨逐渐优化。

此外，在搜索系统中，权重往往被赋予为某一常量，这种设定在某一时期可能会得到良好的搜索结果。然而，随着系统的转型，用户人群口味变更或者源数据的变化等因素，这一设定便将难以获得准确的结果。在多字段检索中，如何根据搜索反馈效果，搜索量等因素动态分配各字段权重以达到最优匹配结果，也是本领域技术人员需要重点考虑的。

例如，搜索系统中的用户一开始只是对某几个主播感兴趣，则他们对主播名的搜索结果更加看重，系统中主播名的搜索量变化加大，搜索反馈效果也会最为良好，权重也会动态偏向这一字段；然而，随着用户对系统的逐渐了解，他们更对房间内容关注更多，则相应的其搜索量便会增大，且反馈效果也更为良好，自然，权重也会偏向到相应的房间名及房间类型。

当系统出现转型或用户搜索喜好发生变化之后，自定义权重变量

随之改变。例如，搜索系统有以下几个字段：主播名、主播房间名、房间类型。系统一开始需要搜索对主播名的搜索更侧重，则只需要增大自定义权重，即字段权重动态分配公式中的自定义权重变量。

本领域的技术人员可以对本发明实施例进行各种修改和变型，倘若这些修改和变型在本发明权利要求及其等同技术的范围之内，则这些修改和变型也在本发明的保护范围之内。

说明书中未详细描述的内容为本领域技术人员公知的现有技术。

权 利 要 求 书

1、一种基于 Lucence 自定义词库的全文检索匹配方法，其特征在于，包括以下步骤：

在基于 Lucence 全文检索引擎的搜索环境中，实时获取用户输入的搜索词，检测是否搜索到结果，如果搜索不到结果，则将搜索不到结果的搜索词去掉特殊字符，存入 Lucence 自定义词库；如果搜索到结果，则对搜索到结果的搜索词进行分词处理，得到分词后的若干词组；对分词后的若干词组继续进行搜索，检测是否搜索到结果，如果搜索不到结果，则将搜索不到结果的分词后的词组去掉特殊字符，存入 Lucence 自定义词库；如果搜索到结果，则记录搜索的时间、分词后的搜索词、搜索反馈信息，最终建立支持 Lucence 全文检索的 Lucence 自定义词库。

2、如权利要求 1 所述的基于 Lucence 自定义词库的全文检索匹配方法，其特征在于：所述建立支持 Lucence 全文检索的 Lucence 自定义词库之后，还包括以下步骤：在建立支持 Lucence 全文检索的 Lucence 自定义词库的基础上，根据字段的搜索量、搜索反馈信息及自定义权重变量线性叠加，按照字段权重动态分配公式定时计算字段权重值，再通过 Lucence 全文检索引擎的权重设置接口，将计算得到的字段权重值动态赋值给字段。

3、如权利要求 2 所述的基于 Lucence 自定义词库的全文检索匹配方法，其特征在于：所述字段权重动态分配公式为：

$$boost = (\alpha * n + \beta * m + \delta * \ln(t) + r) * \rho,$$

其中，boost 为某字段权重值，n 为某时段该字段被检索量，m 为某时段该字段被检索后具有完整搜索反馈的总量，t 为某时段该字段被检索后不完整搜索反馈的总量，r 为自定义权重变量， α 为检索

量的系数因子, β 为完整搜索反馈的系数因子, δ 为不完整搜索反馈的系数因子, ρ 为全局协调系数因子。

4、如权利要求 3 所述的基于 Lucence 自定义词库的全文检索匹配方法, 其特征在于: 所述自定义权重变量为主播名、主播房间名或房间类型。

5、如权利要求 4 所述的基于 Lucence 自定义词库的全文检索匹配方法, 其特征在于: 当系统出现转型或用户搜索喜好发生变化之后, 所述自定义权重变量随之改变。

6、一种基于 Lucence 自定义词库的全文检索匹配系统, 其特征在于: 该系统包括 Lucence 自定义词库建立单元, 所述 Lucence 自定义词库建立单元用于建立支持 Lucence 全文检索的 Lucence 自定义词库: 在基于 Lucence 全文检索引擎的搜索环境中, 实时获取用户输入的搜索词, 检测是否搜索到结果, 如果搜索不到结果, 则将搜索不到结果的搜索词去掉特殊字符, 存入 Lucence 自定义词库; 如果搜索到结果, 则对搜索到结果的搜索词进行分词处理, 得到分词后的若干词组; 对分词后的若干词组继续进行搜索, 检测是否搜索到结果, 如果搜索不到结果, 则将搜索不到结果的分词后的词组去掉特殊字符, 存入 Lucence 自定义词库; 如果搜索到结果, 则记录搜索的时间、分词后的搜索词、搜索反馈信息。

7、如权利要求 6 所述的基于 Lucence 自定义词库的全文检索匹配系统, 其特征在于: 该系统还包括字段权重动态分配单元, 所述字段权重动态分配单元用于动态分配字段权重: 在 Lucence 自定义词库的基础上, 根据字段的搜索量、搜索反馈信息及自定义权重变量线性叠加, 按照字段权重动态分配公式定时计算字段权重值, 再通过 Lucence 全文检索引擎的权重设置接口, 将计算得到的字段权重值动

态赋值给字段。

8、如权利要求 7 所述的基于 Lucence 自定义词库的全文检索匹配系统，其特征在于：所述字段权重动态分配公式为：

$$boost = (\alpha * n + \beta * m + \delta * \ln(t) + r) * \rho,$$

其中，boost 为某字段权重值，n 为某时段该字段被检索量，m 为某时段该字段被检索后具有完整搜索反馈的总量，t 为某时段该字段被检索后不完整搜索反馈的总量，r 为自定义权重变量，例如：主播名、主播房间名、房间类型； α 为检索量的系数因子， β 为完整搜索反馈的系数因子， δ 为不完整搜索反馈的系数因子， ρ 为全局协调系数因子。

9、如权利要求 8 所述的基于 Lucence 自定义词库的全文检索匹配系统，其特征在于：所述自定义权重变量为主播名、主播房间名或房间类型。

10、如权利要求 9 所述的基于 Lucence 自定义词库的全文检索匹配系统，其特征在于：当系统出现转型或用户搜索喜好发生变化之后，所述自定义权重变量随之改变。

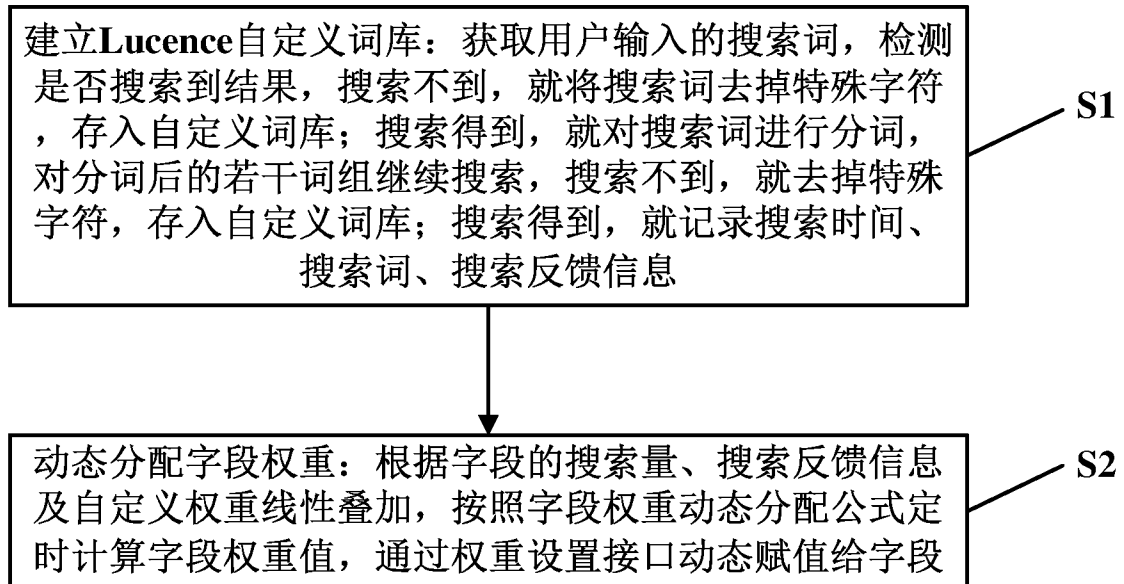


图 1

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2017/080784

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI, EPODOC, CNPAT, CNKI, GOOGLE: the whole document, word segmentation, self-define, Lucence, Lucene, search, query, word, keyword, segmentation, dictionary, stock, lexicon, time, feedback, user, custom, define

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 106021430 A (WUHAN DOUYU NETWORK TECHNOLOGY CO., LTD.), 12 October 2016 (12.10.2016), claims 1-10	1-10
A	CN 103729463 A (CERTUSNET INC.), 16 April 2014 (16.04.2014), description, paragraphs [0002] and [0037]-[0041]	1-10
A	CN 105404661 A (INSPUR (BEIJING) ELECTRONIC INFORMATION INDUSTRY CO., LTD.), 16 March 2016 (16.03.2016), the whole document	1-10
A	CN 102929902 A (JIANGSU XINRUIFENG INFORMATION TECHNOLOGY CO., LTD.), 13 February 2013 (13.02.2013), the whole document	1-10
A	CN 104469832 A (WUHAN HONGXIN COMMUNICATION TECHNOLOGIES CO., LTD.), 25 March 2015 (25.03.2015), the whole document	1-10
A	US 2014280166 A1 (MARITZ HOLDINGS INC.), 18 September 2014 (18.09.2014), the whole document	1-10

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier application or patent but published on or after the international filing date	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&" document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means	
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search
05 July 2017 (05.07.2017)

Date of mailing of the international search report
19 July 2017 (19.07.2017)

Name and mailing address of the ISA/CN:
State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao
Haidian District, Beijing 100088, China
Facsimile No.: (86-10) 62019451

Authorized officer
DAI, Lei
Telephone No.: (86-10) **62414073**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2017/080784

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 106021430 A	12 October 2016	None	
CN 103729463 A	16 April 2014	None	
CN 105404661 A	16 March 2016	None	
CN 102929902 A	13 February 2013	None	
CN 104469832 A	25 March 2015	None	
US 2014280166 A1	18 September 2014	US 2016267169 A1	15 September 2016
		US 9298814 B2	29 March 2016

国际检索报告

国际申请号

PCT/CN2017/080784

A. 主题的分类 G06F 17/30 (2006.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																							
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) WPI, EPODOC, CNPAT, CNKI, GOOGLE: 全文, 检索, 搜索, 查询, 查找, 分词, 词典, 词库, 时间, 反馈, 用户, 自定义, Lucence, Lucene, search, query, word, keyword, segmentation, dictionary, stock, lexicon, time, feedback, user, custom, define																							
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 106021430 A (武汉斗鱼网络科技有限公司) 2016年 10月 12日 (2016 - 10 - 12) 权利要求1-10</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 103729463 A (赛特斯信息科技股份有限公司) 2014年 4月 16日 (2014 - 04 - 16) 说明书第[0002], [0037]-[0041]段</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 105404661 A (浪潮北京电子信息产业有限公司) 2016年 3月 16日 (2016 - 03 - 16) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 102929902 A (江苏新瑞峰信息科技有限公司) 2013年 2月 13日 (2013 - 02 - 13) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 104469832 A (武汉虹信通信技术有限责任公司) 2015年 3月 25日 (2015 - 03 - 25) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>US 2014280166 A1 (MARITZ HOLDINGS INC.) 2014年 9月 18日 (2014 - 09 - 18) 全文</td> <td>1-10</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 106021430 A (武汉斗鱼网络科技有限公司) 2016年 10月 12日 (2016 - 10 - 12) 权利要求1-10	1-10	A	CN 103729463 A (赛特斯信息科技股份有限公司) 2014年 4月 16日 (2014 - 04 - 16) 说明书第[0002], [0037]-[0041]段	1-10	A	CN 105404661 A (浪潮北京电子信息产业有限公司) 2016年 3月 16日 (2016 - 03 - 16) 全文	1-10	A	CN 102929902 A (江苏新瑞峰信息科技有限公司) 2013年 2月 13日 (2013 - 02 - 13) 全文	1-10	A	CN 104469832 A (武汉虹信通信技术有限责任公司) 2015年 3月 25日 (2015 - 03 - 25) 全文	1-10	A	US 2014280166 A1 (MARITZ HOLDINGS INC.) 2014年 9月 18日 (2014 - 09 - 18) 全文	1-10
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																					
PX	CN 106021430 A (武汉斗鱼网络科技有限公司) 2016年 10月 12日 (2016 - 10 - 12) 权利要求1-10	1-10																					
A	CN 103729463 A (赛特斯信息科技股份有限公司) 2014年 4月 16日 (2014 - 04 - 16) 说明书第[0002], [0037]-[0041]段	1-10																					
A	CN 105404661 A (浪潮北京电子信息产业有限公司) 2016年 3月 16日 (2016 - 03 - 16) 全文	1-10																					
A	CN 102929902 A (江苏新瑞峰信息科技有限公司) 2013年 2月 13日 (2013 - 02 - 13) 全文	1-10																					
A	CN 104469832 A (武汉虹信通信技术有限责任公司) 2015年 3月 25日 (2015 - 03 - 25) 全文	1-10																					
A	US 2014280166 A1 (MARITZ HOLDINGS INC.) 2014年 9月 18日 (2014 - 09 - 18) 全文	1-10																					
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																							
<table border="0"> <tr> <td> * 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 </td> <td> “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件 </td> </tr> </table>			* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																			
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																						
国际检索实际完成的日期 2017年 7月 5日		国际检索报告邮寄日期 2017年 7月 19日																					
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		授权官员 戴雷 电话号码 (86-10)62414073																					

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/080784

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	106021430	A	2016年 10月 12日	无			
CN	103729463	A	2014年 4月 16日	无			
CN	105404661	A	2016年 3月 16日	无			
CN	102929902	A	2013年 2月 13日	无			
CN	104469832	A	2015年 3月 25日	无			
US	2014280166	A1	2014年 9月 18日	US	2016267169	A1	2016年 9月 15日
				US	9298814	B2	2016年 3月 29日