文章编号:1000-5862(2015)06-0637-05

云计算环境下可视化探索式搜索引擎的研究

周　莉[1],王　珏[1],周　勇[2]
(1. 华东交通大学软件学院,江西 南昌　330013;2. 江西师范大学计算机信息与工程学院,江西 南昌　330022)

摘要:针对目前搜索引擎返回的信息量过大且缺乏语义关联等问题,提出了一种云计算环境下的可视化探索式搜索引擎模型. 该模型通过对元搜索引擎返回的原始信息在云计算环境下语义相似度的计算和语义链的构建,采用探索式搜索方法为用户获取个性化的结果. 与传统搜索引擎相比,其结果更加直观地表现了目标信息及其之间丰富的语义关系,该方法使用户能够更为自然而有效地在海量的信息中发现更符合其需求的目标. 作为实验模型,还需要更多的元搜索引擎的支持,以及进一步计算优化语义相似度的算法,才能使该模型真正实用化. 本研究为云计算环境下构建新一代个性化智能搜索引擎提供了理论和实践上的参考.

关键词:云计算;探索式搜索;语义关系
中图分类号:TP 393　　**文献标志码**:A　　**DOI**:10.16357/j.cnki.issn1000-5862.2015.06.17

0　引言

搜索引擎是因特网上最为常用的资源之一,它是指根据一定的策略、运用特定的计算机程序从互联网上搜集信息,在对信息进行组织和处理后,为用户提供检索服务,将用户检索相关的信息展示给用户的系统. 作为最流行的商用搜索引擎,谷歌和百度计算搜索结果与查询之间的相关性,并将其以有序列表的形式呈现给用户. 然而,随着信息越来越普及,搜索者对搜索引擎的需求不断增长,不仅仅提供简单的查询行为. 如今,信息搜索已经是同时在人类和信息技术环境中试图获取信息的过程或活动[1],借助搜索进行学习和探索已经变得越来越重要. 然而,现有的搜索引擎对此帮助甚微,大多数的情况下是返回结果过多,以至于用户对其无法充分理解. 此外,普通搜索引擎返回的大量结果之间的关系尚不清楚,因此用户需要投入大量精力找出存在于数量巨大的结果之间的语义关系.

针对上述情况,用户提出了一些应对策略,包括提交多个查询、检索文档空间的互动探索和有选择性地通过链接以被动地获得进一步的线索等[2]. 探索式搜索是一个非常重要的解决方案,它描述用户是如何基于对信息预测值的决策进行搜索,并试图优化其搜索行为. 在进行探索性搜索时,用户所关心的是找到满足其目标的信息,并不是重视找到目标的最优路径,同时,传统的探索性搜索返回的仍然是一个结果列表,其用户体验并不够好.

相比之下,本文提出的搜索引擎框架,能够给出用户感兴趣内容的概况,让搜索者从粗到细地对感兴趣的信息进行浏览. 因此,对所需信息之间关系的图形表示可以显著地帮助用户在搜索中学习和探索. 一般情况下,本文的框架可以有效地同时解决来自用户的 3 个基本需求:(i) 集成和凝练大量来自常用搜索引擎的元搜索结果;(ii) 可视化的探索性搜索,以从海量数据中学习和探索用户的个性化需求;(iii) 友好高效的用户接口和可视化的探索性搜索所带来的可用性和便利性.

1　背景知识

1.1　元搜索

搜索引擎是当今互联网上最成功的应用,其主要功能是帮助用户查找特定主题的信息. 常见的搜索引擎(如谷歌和百度),用户通过搜索文本框提交关键词进行查询并接收文本形式的结果列表. 但不同的搜索引擎使用不同的算法用于收集、索引和搜索信息的链接,因此,对于相同的关键词,不同的搜索引擎可能返回不同的结果列表. 要达到其目的,用户可能需要查询多个搜索引擎. 元搜索引擎通过同

收稿日期:2015-08-17
基金项目:国家自然科学基金(F030408,F020106)资助项目.
作者简介:周　莉(1977 -),女,江西南昌人,讲师,主要从事云计算搜索引擎方面的研究.

时提交一个查询至多个搜索引擎自动完成这一过程,从而减轻了用户的负担[3].

元搜索为探索性搜索提供足够丰富的信息资源. 在其搜索框架中,本文使用谷歌、百度和一个专用搜索引擎作为元搜索引擎的底层搜索引擎,称为成员搜索引擎,见图 1. 在所有的成员搜索引擎获得了返回结果之后,元搜索引擎将结果合并为一个有序列表. 目前大多数搜索引擎将其检索结果表示为信息检索记录集(SRR)呈现给用户. 一个典型的 SRR 由 URL、标题和检索文献的摘要组成[4]. 因此,与 SRR 有关的内容可以用于对来自不同搜索引擎的检索结果进行合并和排序.

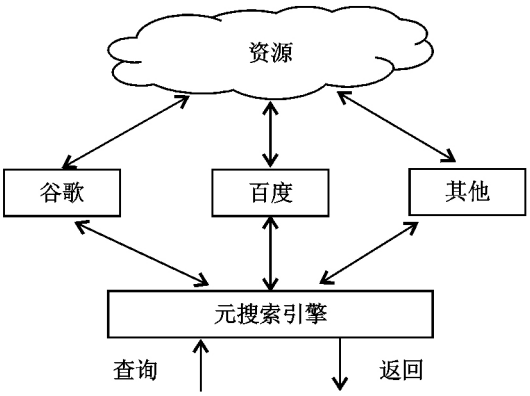

图 1 元搜索引擎

在元搜索引擎中,选择排在列表前列的 SRR,并计算搜索引擎评分的算法称为 TopSRR[4]. 当查询 Q 被提交给搜索引擎 j 时,那么搜索引擎返回动态生成的结果页面,页面上包括一定数目的 SRR,在 TopSRR 算法中,从每个搜索引擎返回的不是前 n 个文档,而是前 n 个文档的 SRR,其用于估算搜索引擎的得分. 直观地说,这是合理的,因为对于一个给定的查询,更好的搜索引擎易于获取更好的结果,结果的优劣通常是反映在其 SRR 之上. 具体而言,来自搜索引擎 j 的前 n 个 SRR 的标题合并在一起,形成一个标题向量 TV_j,所有的片段也被合并成一个片段向量 SV_j. 分别计算查询 Q 和 TV_j 之间的相似度,以及 Q 和 SV_j 之间的相似度,然后汇总到搜索引擎的 j 的评分之上.

算法可以描述为 $S_j = c_1 \times Similarity(Q, TV_j) + (1 - c_1) \times Similarity(Q, SV_j)$,在本文定义的框架中,$c_1 = 0.5$.

1.2 探索性搜索

近年来,出现了一种被称为探索性搜索[1] 的新的搜索方法,探索性搜索是一种特定的信息搜索行为,其搜索者具有以下特征: 不熟悉其目标领域; 实现目标的方法不明确; 目标不明确.

Marchionini 教授将搜索行为分为 3 种主要类型: 查找(即分析搜索从特定查询到其精确结果之间的策略)、学习(即搜索对新知识的认知加工和解释)和调查(即搜索集成至知识库之前所需的关键评价). 他认为是后两个活动构成了探索性搜索[5].

信息检索领域正逐渐加深与人机交互领域的合作,以新的方式使用户更加主动积极地介入搜索过程. 参考文献中提出了几个探索性搜索的原型系统,例如,O. Alonso 等[6] 提出了使用时间轴数据,从而使搜索结果的表示和导览更为有效的新型接口; K. P. Yee等[7] 开发了另一种接口,用于探索使用多层分面元数据和动态生成查询预览的大型图像集合. 然而,上述的探索性搜索原型系统并没有针对不同终端设备上的用户行为进行专门的设计,同时其结果不够直观. 作为一种面向搜索者的方法,友好的用户体验应该作为重要的一部分被纳入到探索性搜索之中.

因此,面向多终端的可视化探索性搜索不仅能够有效地满足用户的搜索需求,同时能够充分利用设备的多样性,体现其无处不在的可用性和便携性等优势.

1.3 语义网络

哲学家和心理学家们早就认识到,人类的记忆是具有联想性的. 联想记忆的第一个计算机模型是 Quillian 的语义记忆系统[8],它基于一个由概念之间的相互联系定义的语义网络.

语义网络用来表达复杂的概念及其之间的相互关系,是一个有向图,其顶点表示概念,而弧则表示这些概念间的语义关系,从而形成一个语义网络描述图[9]. 语义网络的计算机实现首先用于人工智能和机器翻译,可以用来表示知识或支撑知识推理的自动化系统. 近来,信息自动提取和基于文本语料库关系的语义网络的建设成为了一个热门的专业领域,并有着广泛的应用. 目前的语义网络系统分为以下 3 组: (i) 基于单词分布属性的系统: 对词语共生分布进行研究,来计算这些词语所代表的概念之间的语义距离[10]; (ii) 基于模式抽取和匹配的系统: 基于词汇模式或词汇语义模式来发现无限制文本中概念之间的关系,这种关系是本体的和无分类的[11]; (iii) 基于字典定义分析的系统,它利用字典的特殊结构,以提取用于整理本体概念的上下义关系. 概念的定义和注释非常有用,因为它们描述精炼且包含最显著的信息[12]. 本文采用的方法属于第 1 组.

1.4 云服务器

因为该框架面向不同的终端,其计算能力各不

相同,所以将服务器置于云计算平台中,称之为云服务器. 从云服务器中获得的搜索结果在被传递到终端之前,需要进一步的后期处理,语义关系图的预处理和后期处理步骤需要耗费大量的计算资源,终端的计算能力目前无法处理这些计算密集型的步骤. 因此,将这些计算过程提交至云服务器,凭借其强大的计算能力和云计算服务的高可扩展性,搜索引擎框架将实现实时处理.

2　框架

搜索引擎框架的基础架构如图 2 所示. 客户端使得用户以直观交互的方式探索和发现针对特定起始关键词的信息,云服务器通过元搜索和知识集成来构造并返回关系图. 该框架由 3 个主要部分组成: 基于元搜索的语义关系图的推理和表示、对基于查询和浏览策略形成的关系图的可视化探索性搜索,以及人机交互.

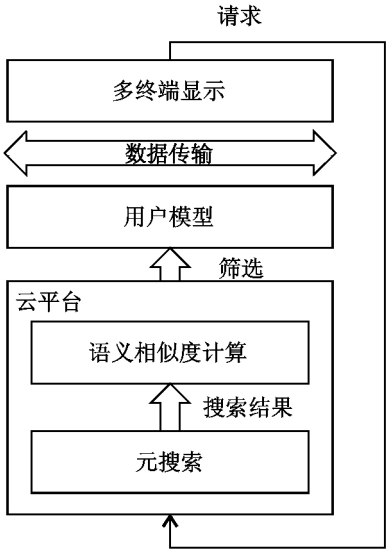

图 2　框架的基础架构

2.1　构造语义关系图

语义关系图的构造包括 3 个主要步骤: (i) 输入关键词; (ii) 在主流搜索引擎上进行元搜索; (iii) 对返回的网页和网站的内在关系进行推理.

在(i) 中,用户输入待查找的关键词; 在(ii) 中,关键词会被发送到一个内部元搜索引擎,该引擎与包括百度和谷歌在内的成员搜索引擎进行交互,如图 1 所示,成员搜索引擎接受关键词并返回检索的链接和网页结果列表. 具体来说,元搜索引擎调用成员搜索引擎的应用程序接口(API); 在(iii) 中,对返回的网页和网站进行有效的自然语言处理和统计,并构建语义关系图. 尤为重要的是,对于所有从元搜索返回的结果,测量其语义相似度,为可视化探索性搜索做好准备. 在所建立的关系图中,各节点表示为关键词,弧被定义为语义关系的长度. 特别的是,返回的结果列表按照图中的弧进行索引.

2.2　计算语义相似度

开发网络搜索机制需要解决两个核心问题: (i) 如何找到相关的网页,(ii) 给定一组潜在的相关网页,如何根据相关性排名. 为了评估网络搜索机制在搜索和结果排名中的有效性,需要进行语义相似度的计算.

从网页中的文本和链接中自动提取语义信息是提高搜索结果质量的关键. 从语言学研究的发展中受益,有较多的字典资源可用于计算语义相似度. 对于英文文本,可以使用 WordNet 作为字典资源,对于中文文本,可用 How-Net. 这些词典资源可被用于计算单词之间的语义相似度,就能得到文件或网页之间的相似度.

文献[13] 介绍了一种基于 How-Net 的语义相似度计算方法. 这种方法将主要、次要和其他关系特征量化,并且它可以减少将补充义义素作为次要特征中的基本义素的错误. 实验结果表明,计算结果向两端扩散,变得更加合理,从而能更准确地区分不同单词之间的微小差异. 本文中框架假定任何 2 个单词(w_1 和 w_2) 的相似度已知,记为 $WordSim(w_1, w_2)$.

假定 2 个文件表示为: $D_1 = \{w_{1,1}, w_{1,2}, \cdots, w_{1,len1}\}$, $D_2 = \{w_{2,1}, w_{2,2}, \cdots, w_{2,len2}\}$, 其中 $len1$ 表示文件 D_1 中单词的数目,同样的,$len2$ 是文档 D_2 中单词的个数. 可以使用 D_1 和 D_2 中目标单词的相似度来测量 D_1 和 D_2 的相似度. 求得目标单词集合 P_t 的算法描述如下: (i) 初始化目标单词集合 P_t 为空集,初始化候选单词集合 P_c 为 D_1 和 D_2 中所有单词的合集; (ii) 找出 P_c 中相似度最大的单词对 $p = (w_i, w_j)$; (iii) 将 p 加入 P_t,从 P_c 中删除所有包含 w_i 或 w_j 的单词对; (iv) 重复步骤(ii) 和(iii),直到 P_c 为空集.

获得目标单词集合 P_t 后,再用以下的方法来计算 D_1 和 D_2 的语义相似度: $SS(D_1, D_2) = 2 \times \sum_{(wi,wj) \in Pt} WordSim(w_i, w_j) / (len1 + len2)$.

用基于 How-Net 的单词语义相似度计算方法计算 $WordSim(w_i, w_j)$,为了得到文件之间更为准确的相似度,在文件中加入单词的权值,并基于单词的权值采用下列方法计算文件的相似度: $W\_SS(D_1, D_2) = \delta / (\sum_{w \in D1} W_{w,d1} + \sum_{w \in d2} W_{w,d2})$, 其中 $\delta = \sum_{(wi,wj) \in Pt} (W_{wi,D1} + W_{wj,D2}) \times WordSim(w_i, w_j)$. $W_{wi,D1}$ 代表单词 w_i 在文件 D_1 中的权值.

2.3 关系图的可视化探索性搜索

一旦关键词有效,就可以通过上述方法推导出描述其关联的语义关系图,其大小由图中几十甚至上百万条的弧来决定. 然后,对这些关系图进行有效甚至是高效的可视化探索性搜索就成为了一个重要的研究目标. 因此,本文提出了一种个性化和智能化的搜索解决方案,通过统计和分析用户的行为,然后自动优化一种最适合用户的独特的搜索模式. 该解决方案将语义关系图的大小最小化.

对于每个用户,需要配置其个性化参数. 系统为用户提供了一些选项. 例如,喜欢体育新闻的用户,他可以直接使用运动类的参数模板. 完成个性化参数的配置后,系统还会根据用户的搜索行为自动优化这些参数. 建议使用开源搜索引擎 Lucene 来搜索该项目参与者的本地个人电脑中的文件. 然后,从用户的个人计算机中获得高频单词,将其组织成为用户的配置文件. 对于已注册的用户,可跟踪其使用历史,征求其输入关键词并来构建用户配置文件. 总而言之,可视化探索的过程从用户输入的关键词开始,然后框架将定位用户感兴趣的节点. 此外,通过图形建模的个性化用户配置文件将通过图形匹配的方法被用于定义整体关系图的子图.

从概念上讲,此基础架构有两个优势. 首先,用户的可视化探索空间可以通过个性化用户模型被相对精确地定位,以便提供给用户最为相关的信息. 因此,用户不需要去探索不相关的搜索空间,这将显著提高搜索信息的用户体验; 其次,不同的配置文件使得个人用户能够探索根据其需要个性化的搜索空间.

实验证明,用户根据其个人兴趣和需要积极参与交互式可视化探索性搜索,从中学习和发现相关信息,这是本文中框架的一大优势.

2.4 框架的业务流程

该搜索系统能够在个人电脑和移动平台上良好运行. 因为其友好的用户界面、无处不在的互联网连接、基于图形的可视化探索性搜索系统尤为适合移动平台. 在本节中,使用智能手机为例来演示该框架. 智能手机的界面是这个框架面向用户的前端,它与云服务器通过无移动网络传输语义关系图和查询结果. 实际上,来自用户的交互事件将由云服务器及时响应,同时,智能手机上的图形进行相应的改变. 另外,从云服务器获得的搜索结果需要进行进一步的后期处理之后才能交付给作为客户端的智能手机,并进行显示. 用户交互的过程如下:(i) 用户在智能手机开始查询,在搜索系统的移动前端上输入关键词,通过互联网与云服务器交互. (ii) 搜索系统将元搜索的结果传输至结果处理系统(RPS). 在此过程中按照本文提出的方法计算网页链接或文件的语义关系. (iii) 用图来表示这些关系. 但是,语义关系图的大小可以包括几百到上百万条弧,由于其太过复杂,难以在终端的屏幕上显示. 因此,采用图匹配的方法,使用根据关系图建模得到的用户配置文件在总体知识图中创建子图. 因此,兴趣图(GOI) 被用于描述特定用户的可视化搜索.

在这个意义上,用户可视化探索的兴趣图 GOI 的初始化可以在图匹配的过程中完成. 出于对移动终端(如智能手机) 有限的显示能力的考虑,原始的 GOI 必须由云服务器按照智能手机的显示与交互能力进行定制. 云服务器首先根据图的层次属性和节点或弧的密度将图中的节点分组,并生成另一个抽象图. 这一步的目的是针对智能手机显示的不同比例和不同分辨率来定制 GOI; RPS 的第 2 步是图的布局. 一个优秀的布局风格会带来良好的用户体验. 布局的目标是建立一个易操作的图并生成其属性用于后续处理; RPS 以查询关键词作为 GOI 结果的中心并以智能手机的分辨率作为约束对图进行筛选; RPS 的最后生成代表 GOI 的 XML 数据流,并将其发送到智能手机; 智能手机接收 XML 数据,解析后显示给用户.

从概念上讲,作为匹配基础框架的图,GOI 有两个主要的优点:(i) 根据用户模型的约束,用户的可视化探索空间可以较准确地定位,因此,可以提供给用户关联程度最高的信息,这将大大减少搜索时间; (ii) 有着各自配置文件的不同用户可以浏览按照其需求定制的探索空间. 例如,即使用户 1 和用户 2 输入相同的关键词,例如苹果,由于其个人配置文件的不同,所以其子图是不同的. 关注个人计算机的用户指的是一台 Mac 电脑,而关心食物用户则指的是一种水果. 因此,这种方法可以有效地实现个性化搜索的目的.

在获得 XML 数据流之后,语义关系图将显示在智能手机上,用节点代表相关的关键词,关键词之间的语义关系的强度则用弧的长度来表示,关键词之间的弧的长度越短,则语义关系越强.

为了方便用户进行可视化探索性搜索,基于知识空间,语义关系图被分为不同的子图类别,并用不同的颜色标示. 用户可以将其个人配置文件与 GOI 进行匹配,以筛选搜索结果,从而为用户最感兴趣的信息提供一个最佳的视图. 如果用户已经找到查询的确切结果,或者找到了需要更多的细节的内容,用户将得到一个独立的网页. 值得注意的是,用户的浏览历史可以用于个性化用户模型构建.

总之,该框架适用于多种终端,尤其是移动设备. 为用户基于语义关系图执行可视化探索性搜索提供了一种新的有效途径.

3　结论和讨论

本文提出的基于关系的探索性搜索引擎模型相比于传统的、基于关键词的查找性搜索引擎,在多个方面,包括丰富的语义、学习和发现、相关性、个性化和更为自然和用户交互等,都存在明显优势.

此外,使用关系来表示诸如 WordNet 中语义的自然语言处理领域,许多学科都使用了交互关系来表示语义和功能的意义. 在网页浏览领域,提出了关系浏览器的概念,用于帮助人们探索不同属性集之间的关系,从而能够知道语料库的范围和程度. 因此,域中各项目之间的关系可能是一个通用的语义表征的方法. 因此,有关基于关系的搜索方法的研究能够促进网络搜索引擎相关学科的发展.

4　参考文献

[1] White R W, Roth R A. Exploratory search: beyond the query-response paradigm [J]. Synthesis Lectures on Information Concepts, Retrieval, and Services, 2009, 1(1): 1-98.

[2] White R W, Drucker S M, Marchionini G, et al. Exploratory search and HCI: designing and evaluating interfaces to support exploratory search interaction [C]. New York: ACM, 2007: 2877-2880.

[3] 安和平,雷英杰,杜书华,等. 元搜索引擎研究 [J]. 计算机工程与设计,2010,31(22): 4787-4789.

[4] 种梅,刘方爱. 元搜索引擎中的成员选择和结果合并策略研究 [J]. 计算机工程与设计,2007,28(21): 5125-5127.

[5] Marchionini G. Exploratory search: from finding to understanding [J]. Communications of the ACM, 2006, 49(4): 41-46.

[6] Alonso O, Baeza-Yates R, Gertz M. Exploratory search using timelines [EB/OL]. [2013-09-14]. http://www.researchgate.net/publication/228826308_Exploratory_search_using_timelines?ev=prf_pub.

[7] Yee K P, Swearingen K, Li K, et al. Faceted metadata for image search and browsing [C]. New York: ACM, 2003: 401-408.

[8] 祝玉芳,王黎华,丁树良,等. 多策略的多级评分认知诊断方法的开发 [J]. 江西师范大学学报: 自然科学版, 2015,39(4): 371-376.

[9] 李蕾,郭祥昊. 基于语义网络的概念检索研究与实现 [J]. 情报学报,2000,19(5): 525-531.

[10] 曾道建,来斯惟,张元哲,等. 面向非结构化文本的开放式实体属性抽取 [J]. 江西师范大学学报: 自然科学版,2013,37(3): 279-283.

[11] 石静,吴云芳,邱立坤,等. 基于大规模语料库的汉语词义相似度计算方法 [J]. 中文信息学报,2013,27(1): 1-6.

[12] 钟伟金. 基于概念关联的词汇语义关系识别研究 [J]. 情报杂志,2014,33(1): 157-160.

[13] 游彬,严岳松,孙英阁,等. 基于 HowNet 的信息量计算语义相似度算法 [J]. 计算机系统应用,2013,22(1): 129-133.

The Research on Visual Exploratory Search Engine in Cloud Computing Environment

ZHOU Li[1], WANG Jue[1], ZHOU Yong[2]

(1. Software School, East China Jiaotong University, Nanchang Jiangxi 330013, China,
2. College of Computer and Information Engineering, Jiangxi Normal University, Nanchang Jiangxi 330022, China)

Abstract: A visual exploratory search engine model based on cloud computing has been proposed for the enormous numbers information and lack of Semantics relationship of the traditional search engine. The model obtains personalized result for users using exploratory search method, via semantic similarity computation and semantic link construction on the raw information returned from meta-search engine. Compared with the traditional search engine, the results are more intuitive on the represent of the target information and their rich semantic relations. Users can more naturally and efficiently found their target in line with their needs in the massive amount of information. As an experimental model, in order to make this model practical, it needs more meta-search engine's support, and further optimization of semantic similarity computation algorithm. The research of the model provides theoretical and practical reference for the construction of new generation of personalized intelligent search engine on the cloud computing environment.

Key words: cloud computing; exploratory search; semantics relationship　　(责任编辑: 冉小晓)