大数据时代的互联网分析引擎

窦志成，文继荣
中国人民大学信息学院大数据管理与分析方法研究北京市重点实验室 北京 100872

摘要
随着互联网尤其是移动互联网的高速发展，互联网文档的数量、内容的丰富度和复杂度都大大增加，互联网正朝大数据时代迈进，而用户的信息需求也趋于复杂化。除了基本的信息检索需求外，对大量相关文档的深入理解与聚合分析的需求也越来越强烈，而传统的互联网搜索引擎已经无法满足人们对该类信息的需求。针对这一问题，提出“互联网分析引擎”的构想，阐述了其与搜索引擎和OLAP分析系统的区别，介绍了一种互联网分析引擎的架构，并详细讨论了实现该引擎的核心问题。

关键词
互联网大数据；分析引擎；数据感知与采集

doi: 10.11959/j.issn.2096-0271.2015027

Web Analytical Engine in the Big Data Era

Dou Zhicheng, Wen Jirong
School of Information & Beijing Key Laboratory of Big Data Management and Analysis Methods, Renmin University of China, Beijing 100872, China

Abstract
Web search engines can only return a list of Web documents (the so-called ten blue links), whereas users may need high-order knowledge that is contained within the Web documents. The demand of analytical services atop the Web is becoming stronger with the rapid development of the internet and the increase of big Web data. The concept of “Web Analytical Engine”, which aims to provide analytical service atop the huge amount of Web documents, was introduced. A simple infrastructure was described and the key research problems for building such an engine were discussed.

Key words
big Web data, analytical engine, data sensing and gathering

1 引言

随着移动互联网、智能手机、社交媒体、自媒体技术的飞速发展以及“互联网+”战略的推广，互联网对个人生活方式的影响进一步深化。互联网从原来仅提供资料发布、信息共享、链接互联等简单功能，开始转变为人们生活中必不可少的信息获取和沟通娱乐的工具，并且发展为与医疗、教育、交通等公用服务深度融合的民生服务。海量的普通用户也逐渐从信息的浏览者变成了信息的制造者。中国互联网络信息中心（CNNIC）发布的《第36次中国互联网络发展状况统计报告》显示，截至2015年6月，中国网站总数为357万个，网民规模达6.68亿户，手机网民规模达5.94亿户[1]。互联网上的信息呈现几何级的增长，体量大、变化快、复杂多样，呈现出典型的大数据特征。

互联网大数据的飞速增长给人们生活带来便利的同时，也导致“信息过载”问题日趋严重。例如，2014年3月8日“马航失联”事件发生后，截至2014年5月21日，仅在百度中被索引的相关网页数量就有500多万篇，新浪微博上关于“MH370”的微博有1 580万条，并产生了大量的转发和评论。如此大量的数据和信息往往超过了个人所能接受的范围。首先，用户从如此海量的互联网数据中查找和浏览有用信息变得越来越困难；其次，用户在查找有用信息的同时会遇到大量的冗余信息；此外，用户在海量的文本内容中进行汇总和理解非常困难。信息检索技术和互联网搜索引擎[2]在一定程度上能够解决上述问题。搜索引擎可以帮用户从海量互联网文档中检索到和用户需求关键词相关的文档，并按照相关性高低进行排序。截至2015年6月，中国搜索引擎用户规模达 5.36 亿户，使用率为 80.3%，搜索引擎是中国网民除了即时通信外使用率最高的互联网应用，并成为人们从互联网获取信息的一个必不可少的工具。但是，随着互联网数据的不断增加以及数据类型的日趋复杂，搜索引擎已经不能很好地满足用户对于信息的深入分析与理解的需求。搜索引擎本质上只能够提供基本的检索功能，而用户往往具有高阶知识获取的需求。例如，当用户在搜索引擎中搜索“马航失联”的时候，很有可能不是在寻找某一条特定新闻或网页，而是希望获取对整个事件或最近进展的一个高度浓缩的知识或结论，如了解“马航失联”事件中各个搜救阶段的主要进行地点和负责机构以及它们之间的关联关系。用户在搜索“天津滨海爆炸”时，是需要了解整个事件的起因、损失情况、救援过程、相关企业信息、民众观点等各方面信息。目前，搜索引擎不能满足用户这种对大规模互联网数据的深层次聚合分析的需求。用户只能先通过搜索引擎或其他应用获取相关网页列表，然后逐一阅读每个网页来对相关内容进行理解和汇总，才能总结出这些检索结果中蕴含的高阶知识。这一过程非常耗时耗力，而在互联网大数据时代，用户也不可能逐一阅读所有相关文档。例如在“马航失联”事件上，百度搜索引擎返回的相关文档有500多万篇，超出了普通用户可以阅读的范围。用户迫切需要一种新的能够帮助用户完成复杂分析任务的系统。和互联网搜索引擎提供的“搜索”功能不同，该系统能够对海量互联网大数据进行深入分析，因此称之为“互联网分析引擎”。互联网分析引擎就像一个“超人”，代替普通用户完成对大规模文档的阅读和理解，并对其中所包含的关键信息与知识进行抽取、挖掘和汇总，并最终通过交互式的分析过程让用户对挖掘到的高阶知识进行浏览和

分析，进而为用户决策提供支持。本文将介绍互联网分析引擎设计构架与数据处理流程，并对其中关键研究问题进行详细阐述。

2 互联网分析引擎

分析引擎旨在提供给用户一个基于海量互联网大数据的多维分析服务，而不仅仅是搜索。搜索引擎重点解决“用户需求的信息在哪里”。给定用户查询后，搜索引擎返回网页或网站列表。例如，若用户查询“雾霾”，搜索引擎可返回一系列关于雾霾的网页和新闻。很多情况下，返回的结果并不能直接满足用户的信息需求。用户仍然需要自己浏览、总结和归纳文档中相关信息。而分析引擎试图在满足用户信息需求的方向上迈进一步，除了找到相关结果外，还要重点回答“这些相关信息从统计上有什么特征”。一个简单的分析引擎中查询“雾霾”的部分输出结果示例如图1所示。该分析引擎可返回雾霾成因、雾霾治理、雾霾成分等维度的内容以及它们的重要性，还可返回关于雾霾的机构、地点、人物、话题、事件等维度以及它们在互联网上的热度随时间变化的趋势。分析引擎还允许用户在分析结果上进行交互。例如，用户在分析结果上选择人物“柴静”，则可进一步分析出在雾霾这一问题上，与柴静相关的互联网信息中其他各维度内容的分布情况：相关的最热话题是“穹顶之下”，相关话题的讨论时间范围是2015年2-3月，这个子话题的相关人物还包括陈吉宁等。

2.1 与搜索引擎的对比

互联网分析引擎和现在广泛使用的互联网搜索引擎的功能对比如图2所示。在搜索引擎的处理逻辑中，文档是基本的检索单位。搜索引擎的核心任务是匹配用户查询词q和互联网上存在的文档d，计算它们的相关性，进而筛选出满足用户意图的文档子集，并按照相关性高低进行排序输出。近年来，虽然各大商业搜索引擎也在不断改变和丰富SERP（search result page，搜索结果页面）的内容，例如集成知识图谱搜索的内容，但搜索结果的主体仍然是网页列表。与搜索引擎类似，互联网分析引擎也以查询词为用户需求的基本表达方式，这一方式延续了这一简单的输入方式给用户带来的便利性。但分析引擎打破了搜索引擎的模式。

第一，系统返回的不再是简单的文档列表，而是高阶知识k。这些知识往往不以具体的形式存在于某个特定互联网文档中，必须对大量文档内容进行理解分析和统计后才能得到。

第二，分析引擎额外强调了时间维度。一方面，分析引擎期望对历史所有文档进行统计分析，结果中可明确地对时间维度进行建模和分析，而现在的搜索引擎一般仅对最新版本的网页进行抓取和处理，这往往忽略了时间维度上所隐含的有用信息；另一方面，在分析引擎中，所处理文档的生成时间和查询时间的间隔要尽量小，即强调分析结果的实时性，而普通的搜索引擎对时效性的要求并不高。

第三，传统搜索引擎能够主要通过简单结果列表的方式展示检索结果，而互联网分析引擎的结果展现和用户交互方式更接近数据仓库系统中的OLAP（online analytical processing，在线联机分析处理）系统[3,4]。主要通过折线图、直方图、面积图、堆积图、饼图、多坐标轴图等统计图表的形式对基于文本立方体的分析结果进行展示，并允许用户基于这些图表

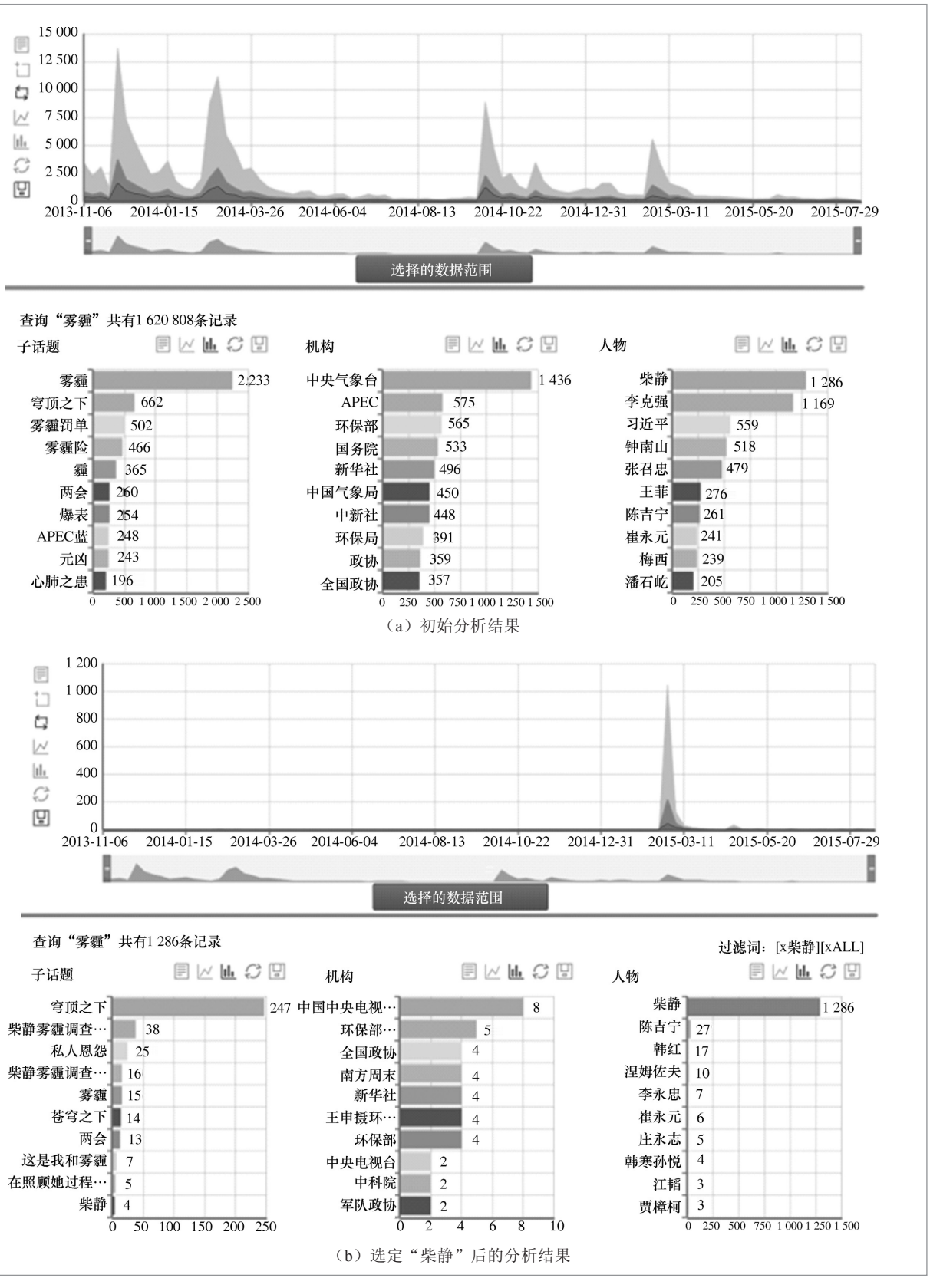


图1 分析引擎中查询“雾霾”的交互式结果示例

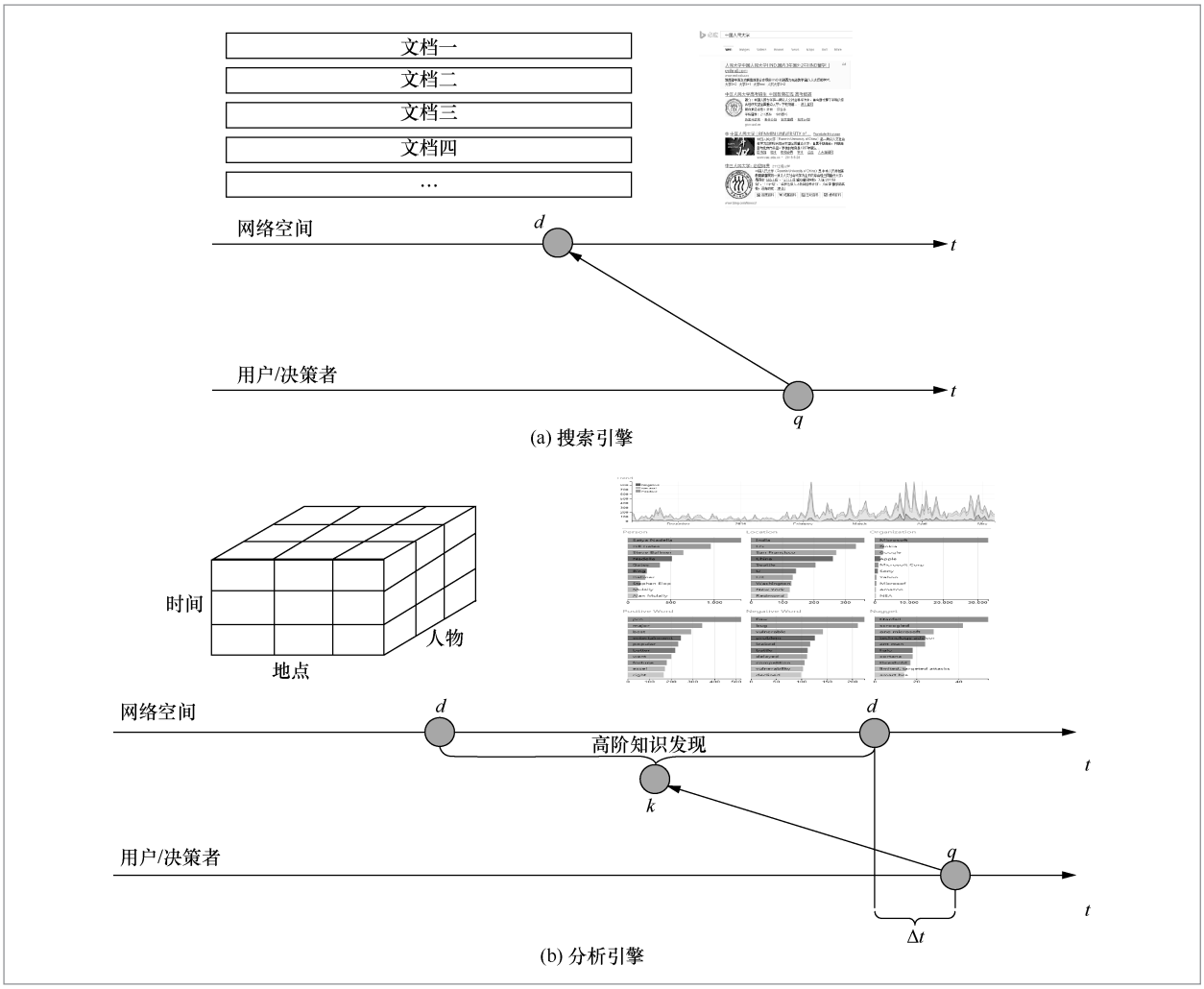


图 2 搜索引擎和分析引擎功能

进行钻取(drill-up和drill-down)、切片(slice)、切块(dice)以及旋转(pivot)等操作,以完成对相关内容的全方位分析。搜索引擎和分析引擎的其他对比见表1。

高效率的检索系统和高质量的检索结果是获得准确分析结果的前提。互联网搜索引擎的底层技术对实现高性能、高质量的分析引擎提供了基础。互联网搜索引擎

表 1 搜索引擎和分析引擎对比

指标	搜索引擎	分析引擎
数据量	大	极大
数据加载处理	批处理	实时、增量
信息处理类型	检索(简单)	分析(文本分析、聚类、统计等)
查询数据量	小	大
查询响应时间	毫秒级	秒级
数据实时性	低	高
并发查询数量	大	较大

主要采用倒排表的方式对文档进行全文索引，并通过对查询词对应的倒排索引列表进行高效的集合操作来匹配文档和查询关键词。搜索引擎在这种高效文档匹配的相关技术上已经非常成熟。给定查询词，搜索引擎通常可以在毫秒级别的时间内从数十亿的互联网文档中匹配到相关文档，选择出相关性最高的前N个结果返回给用户。在这个过程中应用了一系列技术来提高系统性能。例如，通过对文档进行预处理以便进行高效的Top k文档初选[5]，从而大大减少后续操作涉及的文档数量。文档相关性打分等操作仅仅需要在小规模的满足初选条件的文档集上进行，这将大幅度提高检索性能。并且仅仅需要对要显示给用户的某一页文档（通常是10个）生成摘要，而不是对所有匹配文档都生成摘要，也大大节省了磁盘读取和CPU计算代价。而在互联网分析引擎中，搜索引擎中的某些优化策略将不再生效。例如，分析引擎通常需要对匹配到的所有文档进行汇总和计算。如果仅仅在返回的前几个结果上应用分析操作，则很可能因为数据量不足导致分析结果不准确。这意味着分析引擎的I/O和CPU开销将远远高于搜索引擎。

2.2 与OLAP技术的区别和联系

OLAP系统[3,4]是一种基于结构化数据和数据仓库的分析系统，专门用于支持复杂的分析操作，侧重对决策人员和高层管理人员的决策支持。最为常见的方法就是基于多维数据构建数据立方体（cube）模型。通过大量的预聚集计算，生成支持多维分析的数据立方体，并在此基础上支持下钻、上卷、切片、切块、旋转等OLAP操作[6~8]。

OLAP主要构建在结构化数据的基础上，而互联网分析引擎处理的对象则主要是非结构化的互联网文档，如网页、微博、帖子等。与传统的OLAP多维分析技术相比，互联网分析引擎的挑战更大，主要原因如下。

（1）数据无结构。大部分互联网文档是无结构的文本数据，无法像结构化数据一样预定义数据模式（schema），因此处理起来更为复杂。例如在电信业大数据中，基本的通话记录可由主叫号码、被叫号码、通话时长、主叫地点、被叫地点等可枚举的强类型字段构成，并且这些字段的值一般可由数据源直接获取。文本数据一般由自然语言生成，每个无模式的文档记录由不定个数的单字构成，不具有可确定的字段。

（2）依赖于复杂的自然语言理解技术。如前所述，文本数据一般由自然语言生成。如果希望在单字的基础上进一步理解文本包含的语义和知识，例如理解文本包含的实体（人、地点、机构、时间等），则要依赖于复杂的文本挖掘和自然语言理解技术。通过自然语言理解技术获取结构化内容的准确性往往依赖于所采用的分析技术，因此分析引擎中数据中的不确定性因素更多，可疑性（veracity）更高。

（3）开放主题。文本数据的主题和值域是开放的。在传统的分析服务中，每个应用的主题是唯一或者有限的。在整个数据集一般可建立有限个数的数据立方体，通过固定的维度（如区域和时间等）对其进行统计和分析。而互联网数据的主题是开放的。例如，每天的互联网文档可能是在分别描述成千上万个无任何关系的主题，每个主题所涉及的维度和文档可能完全不同，其复杂度远远高于OLAP系统。

近年来也有部分学者开始研究如何将OLAP技术应用在分析大规模互联网数据上。但目前的研究主要针对语义网和RDF

数据[9,10]。如何将OLAP技术应用在大规模互联网文档上来实现互联网分析引擎，仍然是一个未被深入研究和讨论的问题。

2.3 小结

事实上，在数据库和数据挖掘领域，OLAP是为了解决OLTP（online transaction processing，在线事务处理）系统分析处理能力低下的问题而被提出的。在互联网上，搜索引擎相当于一个OLTP系统。用户的每一个查询，搜索引擎都能快速地返回检索结果集。但和OLTP的问题类似，搜索引擎无法有效支持分析处理的需求，而互联网分析引擎也正是为了解决这一问题而生。因此，可以把互联网分析引擎看作互联网搜索引擎和OLAP技术的合体，或者说互联网分析引擎是面向海量互联网非结构化大数据的OLAP系统。

3 互联网分析引擎的设计

如前所述，互联网分析引擎和搜索引擎及OLAP系统都是紧密相关的，在设计分析引擎时，可充分吸收和利用现有系统和算法中的优点，并将二者进行有机结合。简单的想法是先通过搜索引擎技术检索到相关文档，然后利用OLAP分析技术对检索结果进行分析。

一个简单的互联网分析系统架构如图3所示。整个系统分为离线处理和在线处理两个部分。离线部分主要完成数据获取并将文本处理成结构化数据，对结构化数据进行索引。在线处理部分主要完成相关文档检索并基于检索到的结果，对其中包含的结构化知识信息进行高效率的汇总分析操作。

3.1 离线处理

离线部分主要包括互联网数据采集、文档理解及结构化数据抽取、数据索引等几个部分。

（1）数据采集部分与搜索引擎中的数据采集系统类似，使用网络爬虫对互联网内容进行抓取。但互联网分析引擎在数据抓取时还需要考虑抓取周期和抓取策略对最终分析结果的影响，避免因为数据抓取

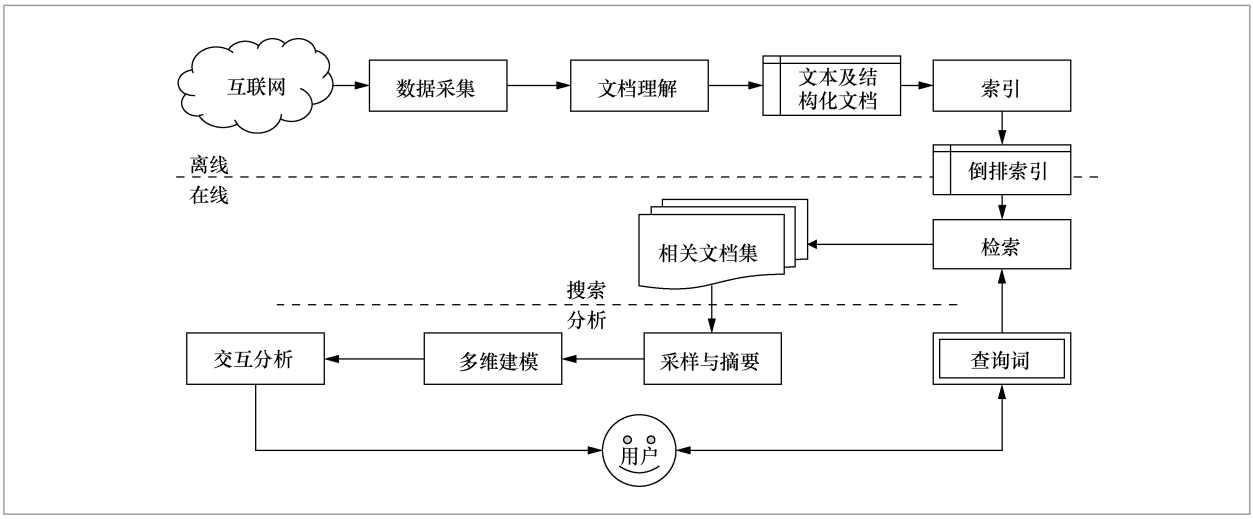


图 3 分析引擎架构

不及时或者数据来源分布不均衡而影响分析结果的一致性和可比性。

（2）文档理解部分主要是利用自然语言处理及信息检索技术，对互联网文档内容进行深入分析，从无结构的文本数据中抽取结构化信息，并将这些结构化数据作为该文档的属性或字段进行存储。将无结构的互联网文档转换成结构化数据后，才能应用OLAP等多维分析技术对文档进行分析。可进行的文本理解工作包括：文档正文及相关属性（标题、时间、作者、主要图片等）抽取、文档内容段落及句子切分、文本分词、命名实体（时间、地点、人物、机构等）识别、动词、专有名词抽取、情感分析及情感词抽取、关键词抽取、引言、语录抽取、知识库实体匹配及消歧等。

（3）数据索引部分对互联网文档内容建立高效索引，以支撑高效的在线检索和分析操作。与搜索引擎类似，在文档内容上建立倒排索引，用以支持基于关键词的文档检索。而对于从文档中抽取出来的结构化属性，则可能既需要建立倒排索引，也需要建立正向索引。倒排索引用于在切片和切块过程中快速匹配筛选维度，而正向索引用于快速获取自定文档的属性。

3.2 在线处理

在线处理部分负责接收用户查询，检索相关文档，并应用OLAP技术，快速检索、构建或更新文本立方体。在整个在线处理流程中，基于信息检索和搜索引擎的积累，检索相关文档所用的时间较短。在进行多维分析时，需要读取到所有相关文档的结构化属性内容，并需要对这些内容按照维度进行汇总和计算操作。与检索相关文档相比，在检索结果上进行多维分析的时间代价要高得多。

4 技术要点与难点

4.1 数据质量控制及可信度评估

提供可靠、可信、有理有据的分析结果是互联网分析引擎能够实用并且推广的前提条件。互联网分析引擎对质量控制和可信度评估的相关技术要求要远远高于搜索引擎。搜索引擎采用了垃圾网页识别技术，尽量减少低质量网页出现的几率，提高用户满意度。但事实上，搜索引擎为用户返回的是在互联网上真实存在的文档（即使是低质量的网页），因此从某种意义上讲，搜索引擎中不存在数据可靠性问题，因为用户需要自己阅读网页内容、判别真伪并形成结论。而在分析引擎中，系统为用户返回的不仅仅是真实存在的网页，而且包括通过加工处理和聚合汇总后的数据，如果这些数据是错误或者有偏差的，则很可能直接导致用户形成错误的结论。因此，在分析引擎中，从数据采集、处理和分析的各个步骤，都需要进行适当的质量控制。例如，在进行数据采集时，适当控制数据采集的广度，避免片面采集某一网站的数据而造成偏差。同时，在各个关键环节需要评估各处理对最终结果可信度的影响。如何在规模巨大、更新飞快、复杂多样的互联网大数据上，针对分析引擎的需求进行质量控制和可信度评估，是非常困难但也非常重要的研究课题。

4.2 大规模文本立方体管理

文本立方体是对某一查询匹配的文档中包含的结构化属性数据进行统计并建立的多维数据立方体。互联网分析引擎

中每个开放的主题或者每个查询都可以建立一个对应的文本立方体。不同的查询都可以有不同维度和度量值的文本立方体，而且可以独立管理。例如，查询“马航失联”对应了一个文本立方体，而“雾霾”则对应了另外一个文本立方体，这两个文本立方体中的数据、维度和值项都可以是不同的。单个文本立方体的规模可能小于传统的数据立方体，但会有大量小规模的文本立方体（many small cube）。这种大量小立方体的分析管理方式最大的优点是灵活，每个查询都可以进行单独的维度和度量值管理，而且对每个小立方体的创建和更新不影响其他立方体。同时，这种设计也便于扩展（scale out），当用户或查询数量增加时，可以简单地增加服务器，并将立方体均匀分布在所有服务器上即可完成系统复杂均衡。LinkedIn公司的Wu等人[11]开发了针对互联网级别OLAP分析的系统Avatara，解决了大量小立方体的问题，可以尝试在互联网分析引擎中应用。

除了创建和管理大量小规模文本立方体外，分析引擎中也可以试图整合所有文本立方体而创建一个超大的通用文本立方体（one giant cube）。该超级立方体中包含所有互联网文档以及所有可能的维度及度量值。这种方式的好处是减少了大量文本立方体管理的代价。这种方式的问题是如果某个查询或某类查询下的分析维度发生变化时，很可能需要重新对整个立方体进行重建操作。当查询之间维度设置差异较大时，在某些查询下进行相关维度的查询和分析的代价可能要高于多个小文本立方体的设置。在系统扩展方面，单个超级立方体的配置下对网络之间的同步以及负载均衡的管理机制更为复杂。

无论是哪种方式，如何高效地进行文本立方体管理都是互联网分析引擎要解决的核心问题，也是难点之一。文本立方体内部存储结构如何设计、如何高效地创建文本立方体、如何动态更新立方体、如何存储和管理大量大规模或大量文本立方体，都是非常重要的研究问题。此外，互联网分析引擎对数据的实时性要求较高，在文本立方体更新和查询操作的同步上也需要仔细斟酌。

4.3 分析维度挖掘与排序

互联网分析引擎的核心目标是为用户提供准确且有效的多维分析结果。除了前文介绍的质量控制和可行度评估外，如何挖掘出高价值的分析维度和度量项、如何对维度中的内容进行排序等也都是需要解决的问题。

在维度发现与挖掘方面，一方面可预设一些通用性的维度，如时间、人物、机构、地点等。同时，还需要在这些基本维度的基础上，挖掘出和用户查询主题相关的个性化维度。例如对于查询“糖尿病”，挖掘出“类型”、“症状”、“药物”、“医院”、“医生”等相关维度；对于查询“过失失火”，可自动挖掘出“刑罚”和“罪名”等维度。只有这样，才能使分析引擎的输出结果变得有用且有趣，才能真正满足用户真实的信息需求。可选的方法是分领域创建维度列表并在离线部分对文档内容和维度列表进行匹配，在线通过分类的方法确定查询所属的领域来获取相关维度。维度的生成可以通过统计分析查询所匹配的文档中包含的属性及结构化数据进行自动聚类和加权，进而自动选出最相关的维度。

在维度及度量项排序方面，在基于OLAP的分析模型下，用于建立文本立方体的每一条数据都需要提供一个度量

值，该度量值决定了在最终文本立方体中每个统计项的权重。和传统的数据立方体（例如基于业务数据生成的立方体）不同，在文本立方体中没有直接的度量值可以使用。文本立方体中的度量值可以通过不同的方法生成，从数据独立性的角度上可以分为下面3种不同类型的度量值。

- 全局一致的度量值。每个文档（记录）的度量值一致，最简单的是每个文档的度量值都为1。
- 与维度值无关的度量值。度量值和记录有关，但和记录中包含的维度无关。例如，考虑到报道的可靠性，所有来自“新浪网”的报道的度量值高于来自“回龙观社区网”的报道的度量值。此外，还可考虑应用信息检索模型来估计文档和主题（查询）的相关性[12~20]，例如，若某个文档和查询的相关性较高，则其度量值较大。
- 和维度相关的度量值。进一步考虑文档（记录）和维度的紧密程度，如对于相关人物A，考虑人物A在文档D中出现的次数、出现的位置、所在句子的长短等特征，并同时考虑报道的来源，从而计算人物A在文档D中的度量值。而对于另一相关人物B，即使同样出现在文档D中，因为人物B的出现次数及位置和人物A不同，人物B的度量值也可能和人物A不同。

和搜索引擎中的搜索结果排序模型一样，在分析引擎中的维度以及度量项排序是非常重要但也是非常复杂的。分析维度挖掘与排序方法是互联网分析引擎要重点研究的问题之一。

4.4 数据采样与摘要技术

因为分析引擎中处理的互联网文档数量非常庞大，而一个查询特别是热点查询往往可以匹配上大量的相关文档。在分析引擎中，匹配文档代价较低，而对相关文档上相应结构化数据的汇总分析和维度生成等操作则具有较高的I/O和计算代价。因此，当数据量太大的时候，在不影响分析结果质量的前提下，可以考虑对匹配到的结果集进行采样、摘要和压缩操作。在数据采样方面，在建立多维模型的时候不能对维度和子主题的优先级进行任何假设，对于任何子主题的数据搜集，都需要保证搜集到足够多的填充数据来体现它的真实语义，力图花费最小的代价重构一个子主题内部的信息点覆盖。可结合维度排序以及维度中包含的值的可信度来估计采用规模。同时，对周期性和长期热点话题采用可合并式数据摘要，并和文本立方体结合，力图通过选择性地保留一部分原始数据和总体上的摘要数据，便能够达到与使用全部数据类似分析效果的目的。同时，结合前文介绍的可信度评估方法，准确计算出各种采用和摘要方法对最终分析效果的影响，力图在系统效率和效果之间达到一个合理的平衡点。

5 结束语

在互联网大数据时代，用户对获取互联网上蕴含的高阶知识的需求也越来越强烈。传统的搜索引擎已经不能很好地满足用户对互联网文档进行深入分析与理解的需求，迫切需要发展到“互联网分析引擎”，为用户提供更为便利的信息获取与分析工具。互联网分析引擎比互联网搜索引擎和OLAP系统更复杂，涉及一系列需要解决的研究难点问题，具有广阔的研究和发展空间。

参考文献

[1] 中国互联网络信息中心. 第36次中国互联网

络发展状况统计报告. http://www.cnnic.net.cn/hlwfzyj/ hlwxzbg/hlwtjbg/201507/P020150723549500667087.pdf, 2006
China Internet Network Information Center. The 36th China Internet Development Report. http://www.cnnic.net.cn/hlwfzyj/ hlwxzbg/hlwtjbg/201507/P020150723549500667087.pdf, 2006

[2] Sergey B, Lawrence P. The anatomy of a large-scale hypertextual Web search engine. Computer Networks, 1998(30): 107~117

[3] Codd E F, Codd S B, Salley C T. Providing OLAP (Online Analytical Processing) to User-Analysts: An IT Mandate. E F Codd &Associates, 1998

[4] Thomsen E. OLAP Solutions: Building Multidimensional Information Systems (2nd Edition). Hoboken: John Wiley & Sons, 2002

[5] Zhu M J, Shi S M, Li M J, *et al*. Effective top-k computation with term-proximity support. Information Processing and Management, 2009(45): 401~412

[6] Gray J, Bosworth A, Layman A, *et al*. Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals. Proceedings of IEEE Computer Society the 12th International Conference on Data Engineering, Washington DC, USA, 1996: 152~159

[7] Han J, Wang J, Dong G, *et al*. Cube explorer: online exploration of data cubes. Proceedings of the 2002 ACM SIGMOD International Conference on Management of data, Madison, Wisconsin, USA, 2002: 626~626

[8] Harinarayan V, Rajaraman A, Ullman J D. Implementing data cubes efficiently. Proceedings of ACM SIGMOD Conference, Montreal, Canada, 1996: 205~216

[9] Etcheverry L, Vaisman A A. Enhancing OLAP analysis with web cubes. Proceedings of the 9th Extended Semantic Web Conference, Heraklion, Crete, Greece, 2012: 469~483

[10] Colazzo D, Goasdou F, Manolescu I, *et al*. RDF analytics: lenses over semantic graphs. Proceedings of the 23rd International Conference on World Wide Web, New York, USA, 2014: 467~478

[11] Wu L L, Sumbaly R, Riccomini C, *et al*. Avatara: OLAP for web-scale analytics products. Proceedings of the VLDB Endowment, Istanbul, Turkey, 2012: 1874~1877

[12] Salton G, Wong A, Yang C S. A vector space model for automatic indexing. Communications of the ACM, 1974(18): 613~620

[13] Croft B, Lafferty J. Language Modeling for Information Retrieval. Norwell: Kluwer Academic Publishers, 2003

[14] Lafferty J, Zhai C X. Probabilistic relevance models based on document and query generation. Language Modeling for Information Retrieval, 2003

[15] Zhai C X, Lafferty J. A study of smoothing methods for language models applied to ad hoc information retrieval. Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieva, New Orleans, Louisiana, USA, 2001: 334~342

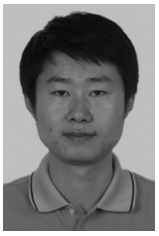
[16] Tao T, Wang X H, Mei Q Z, *et al*. Language model information retrieval with document expansion. Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL'06), Stroudsburg, PA, USA, 2006: 407~414

[17] Srikanth M, Srihari R. Exploiting syntactic structure of queries in a language modeling approach to IR. Proceedings of the 12th International Conference on Information and Knowledge Management, New York, NY, USA, 2003: 476~483

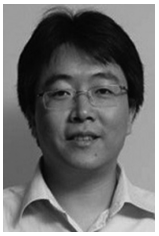
[18] Bai J, Nie J Y, Cao G. Using query contexts in information retrieval. Proceedings of the 30th Annual

International ACM SIGIR Conference, Amsterdam, Holland, 2007: 15~22
[19] Turtle H, Croft W B. Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems, 1991(9): 187~222
[20] Li Z W, Wang B, Li M J, *et al*. A probabilistic model for retrospective news event detection. Proceedings of the 28th Annual International ACM SIGIR Conference, Salvador, Brazil, 2005: 106~113

作者简介



窦志成，男，中国人民大学信息学院研究员、硕士生导师，中国计算机学会大数据专家委员会通讯委员，中文信息学会信息检索专委会通讯委员，中国中文信息学会青年工作委员会委员，亚洲信息检索协会Steering Committee成员，美国ACM学会、IEEE会员，中国计算机学会会员。主要研究方向为信息检索、互联网搜索、数据挖掘、大数据等。近年来，在国际知名会议和学术期刊上（如SIGIR、WWW、CIKM、WSDM、EMNLP及IEEE TKDE等）发表论文20余篇。



文继荣，男，博士，中国人民大学信息学院教授、博士生导师，国家“千人计划”特聘专家。1999年至2013年就职于微软亚洲研究院，自2008年起担任高级研究员和互联网搜索与数据挖掘组主任。在微软亚洲研究院工作的14年中，获得50多项美国专利，其中一些成果已经被用于重要的微软产品中（如微软搜索引擎Bing）。所领导的研究团队开发出了微软学术搜索（http://academic.research.microsoft.com）、人立方（http://renlifang.msra.cn/）、产品搜索等有影响力的互联网应用。在国际著名会议和期刊上发表了100多篇论文，担任过许多国际会议和研讨会的程序委员和主席。目前是信息检索领域主要期刊ACM Transactions on Information Systems（TOIS）的副主编。

收稿日期：2015-08-20
基金项目：国家重点基础研究发展计划（“973”计划）基金资助项目（No.2014CB340403），国家自然科学基金资助项目（No.61502501），中国人民大学科学研究基金（中央高校基本科研业务费专项资金资助）（No. 15XNLF03），国家文化科技提升计划
Foundation Items: The National Basic Research Program of China(973 Program)(No.2014CB340403), The National Natural Science Foundation of China(No. 61502501), The Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China (No. 15XNLF03), The National Culture Science and Technology Promotion Plan
论文引用格式：窦志成，文继荣. 大数据时代的互联网分析引擎. 大数据，2015027
Dou Z C, Wen J R. Web analytical engine in the big data era. Big Data Research, 2015027