doi:10.11959/j.issn.1000-436x.2015317

智慧企业中的智慧搜索

陈扬斌[1]，李青[1]，庄越挺[2]
（1．香港城市大学 电脑科学系多媒体软件工程研究中心，中国 香港 999077；
2．浙江大学 计算机科学与技术学院，浙江 杭州 310027）

摘 要：现代企业除了面临复杂的生产环境和网络环境外，还需积极应对和处理随之产生的海量数据。这些数据服务于企业发展是智慧企业的目的之一。基于企业各个环节可能产生的各种数据类型和对应的搜索技术，智慧搜索旨在为智慧企业的实现与发展增加一种新的智慧服务。通过实例来阐述智慧搜索的内涵和外延，以及智慧搜索能为企业带来的不同级别的服务和相应的挑战。
关键词：智慧企业；智慧搜索；大数据；文本搜索
中图分类号：TP391.3 **文献标识码**：A

Smart search in smart enterprise

CHEN Yang-bin[1], LI Qing[1], ZHUANG Yue-ting[2]
(1. Department of Computer Science, Multimedia Software Engineering Research Center, City University of HongKong, HongKong SAR 999077, China;
2. College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

Abstract: Modern enterprises are facing not only complex production and networked environments, but also massive amount of data generated from the various processes. One of the goals of smart enterprise (SE) is to make full use of the big data for further development of enterprises. By examining the types and applicable search techniques of the data which may come from different process of the enterprise. The notion of smart search was introduced, as a new mechanism to facilitate the implementation and development of SE. Smart search is elaborated through real-life examples, and discuss how smart search can bring different levels of services as well as the challenging issues to be tackled.
Key words: smart enterprise; smart search; big data; text retrieval

1 引言

著名企业家 Lonsdale 在 2013 年的《智慧企业浪潮》一文中，将智慧企业列为硅谷技术趋势里的第 6 个最新浪潮。虽然目前业界对智慧企业的概念尚无明确的定义，但是一个比较普遍被认可的观点是，智慧企业需要满足 3 种特性：1) 各种类型大数据的集成，以帮助知识工作者解决非线性问题；2) 消费浪潮所驱动的信息技术应用到大型工业中；3) 凭借网络效应在垂直行业中能够成为平台，并让新技术在行业中快速传播从而提高整体创新力。本文中，智慧企业所指的“智慧”是企业能够在复杂的数据环境下，综合运用云计算、机器学习、数据挖掘等技术帮助分析、预测、解决问题的一种商业智能。

智慧企业的基础是数据，因此对数据的深入挖掘和应用是其区别于传统企业的重要特点之一。传统的信息检索通过“查询—文档”相匹配的搜索模式，从大数据集合中找到所需的非结构化资源[1]。而智慧搜索（简称“智搜”）是“任务驱动型”的综合性搜索方法。它的输入是一项任务 T，期望输出是一套解决方案 P；任务 T 经逐级分解，通过节点搜索和结果集成等步骤最终返回解决 P 的方案。而整个过程中往往还需要引入领域知识。在数据容量大、可靠性低、增速飞快、类型多样的背景下，

收稿日期：2015-10-11；**修回日期**：2015-12-10
基金项目：国家自然科学面上基金资助项目(61472337)；科技部国际科技合作专项基金资助项目(2014DFG12370)
Foundation Items: The National Natural Science Foundation of China(61472337); The International S&T Cooperation Program of China (2014DFG12370)

智慧搜索更注重结果的知识性，其目的是在企业层面能给企业决策提供一种新的功能途径。

本文在列举企业环节产生的不同数据类型和相应搜索技术的基础上，重点介绍“智搜”的理念与机制，目的是为智慧企业的实现与发展增加一种新的知识服务。

2 基于数据类型的搜索模式

2.1 企业数据类型

无论大、中、小企业，与其业务相关的数据随着企业信息化水平的提升都在不断增加。企业的数据增长来自 3 个方面：1) 数据的迁移，比如办公电子化和建立云平台，实现了数据载体由纸张到计算机的变革；2) 数据采集能力的提升，通过新型传感器技术将生产过程中的数据更为详细和精确地记录下来；3) 新媒体的发展所带来的数据量激增，使企业在市场运营这一块的形式更多种多样，而且数据呈现出多源、多模态的特点。

企业的管理、生产、销售、研发等环节会产生不同类型的数据。这些数据可按其类型分为结构化数据、非结构化数据和半结构化数据。结构化数据以表的形式存储在数据库中，具有确定属性和格式，如财务报表、产品销售记录等。非结构化数据无固定格式，涵盖文本、音频、视频、图片等类型，如产品说明书、各类新闻及评论等。半结构化数据介于结构化数据和非结构化数据之间，通常指 XML、HTML、JSON 等既包含结构化又包含非结构化内容的数据，比如网页数据等。表 1 列出企业中不同部门产生的数据及其所属类型。

2.2 相关搜索模式

企业数据的特点决定了搜索的模式。站在整个企业生态圈的角度，数据搜索的范围可以从企业内部扩大到企业外部的公共资源，也可以跨行业跨平台，数据搜索的方式可以有纵向搜索、横向搜索和“泛”搜索。纵向搜索是从行业细分的角度来搜索产品上下游信息，如生产电器零部件的企业要制定某季度的生产计划，除了依据订单外，历年同期的生产与销售数据、供应链上游原材料厂商的数据、供应链下游电器厂商的数据等都可以作为参考。横向搜索是从同业竞争的角度，在遵守行业规则的前提下尽可能多地获取本行业的情报，以服务于本企业的决策制定。纵向搜索和横向搜索都是在企业内部或企业上下游之间执行。“泛”搜索则着眼于大环境的数据，尤其是 Web 2.0 时代诸如电子商务、社交网络、互联网媒体等平台所产生的开放式大数据。随着信息传播方式的变化和多屏互联网设备的普及，企业服务的主体——客户与企业之间互动越来越多。电商平台可以将产品销售情况、客户评论信息回馈给企业，企业自身可通过社交网络和新闻平台检索出当下大众的关注热点和审美趣味，从而快速地做好产品设计和制定营销模式。高质量的数据是企业的宝贵财富，有助于企业发展。面对日益增长的数据，搜索什么、如何搜索是现代企业应当考虑的重要问题，更是智慧企业需要迎接的一大挑战。

3 搜索技术和算法

3.1 结构化数据搜索技术

企业中的结构化数据一般就是指存储在关系型数据库中的数据，因此对这类数据的搜索使用结构化查询语言即可。随着数据量的增加，关系型数据库的研究重点在于提升存储能力。

3.2 半结构化数据搜索技术

半结构化数据通常是 XML 数据，对 XML 文件的搜索技术包括有 LCA[2] 及其衍生算法如 SLCA[3] 和 VLCA[4] 等。

3.3 非结构化数据搜索算法

非结构化数据包括文本和多媒体数据。下面将主要概括长文本和短文本常用的搜索算法，最后简要介绍一些多媒体检索方法。

3.3.1 长文本

文本搜索是计算查询与待搜索文档间的相关

表 1 企业可能产生的各类文件及其数据类型

数据类型	企业管理	生产过程	公司运营	技术研发
结构化数据	财务报表	原料使用记录， 监测数据记录	产品销售记录， 用户信息记录	数据库文件
半结构化数据	人事档案	生产日志	网站数据（XML, HTML）	需求分析文档，软件设计报告
非结构化数据	文本及多媒体文件	质检报告，产品说明书	用户评论，社交媒体，网络媒体	用户使用指南，专利说明书

度并返回相关度值高的文档的过程。当前主要的搜索模型有向量空间模型[5,6]、概率模型[7,8]和概率语言模型[9,10]。

向量空间模型将查询与文档都转化成基于标引项(term)的向量，二者的相似度常用余弦函数衡量。向量空间模型的研究重点在于如何确定每一维向量值。有 3 个影响因子，它们分别是：1) 标引项在文档中的频率；2) 包含该标引项的文档在文档集合中的频率；3) 文档的长度。除了经典的 TF*IDF 模型[11]外，在向量空间搜索模型中，临界点归一化检索模型是效果最好的模型之一[12]。基本原理如下：用 D 表示文档，Q 表示查询，t 表示一个标引项，S 表示搜索函数，则表示文档与查询的相关度 $S(Q, D)$可由以下公式来计算

$$S(Q,D)=\sum_{t\in D\cap Q}\frac{1+\ln(1+\ln(c(t,D)))}{(1-s)+s\frac{|D|}{avdl}}c(t,Q)\ln\frac{N+1}{df(t)}$$

其中，$c(t, D)$表示标引项在文档中的数量，$c(t, Q)$ 表示标引项在查询中的数量，$|D|$表示文档长度，$avdl$ 表示文档平均长度，N 表示文档集合中文档的总数量，$df(t)$ 表示包含标引项的文档总数量，s 是参数。向量空间模型的不足主要是因为标引项的选择过于粗糙导致向量维度过大，标引项各自独立的前提在现实中不成立以及向量空间模型无法解决语言中经常出现的一词多义或者同义词现象等。

概率模型是文本搜索另一个重要的模型，主要通过估计文档与查询相关的概率返回搜索结果。Okapi BM 25 是性能较好的概率搜索模型[13]，其计算公式如下

$$S(Q,D)=\sum_{t\in D\cap Q}\ln\frac{N-df(t)+0.5}{df(t)+0.5}\cdot\frac{(k_1+1)c(t,D)}{k_1((1-b)+b\frac{|D|}{avdl})+c(t,D)}\cdot\frac{(k_3+1)c(t,Q)}{k_3+c(t,Q)}$$

其中，k_1、b_2、k_3 是参数。

概率语言模型则是另一种被广泛应用的搜索模型。随着自然语言处理技术的发展，统计语言建模技术已逐渐成为当前语言信息处理的主流技术之一[14]。语音识别、机器翻译等研究都用它作为基本模型，而且在文本搜索方面的应用也得到了较好的结果。在文本搜索中，针对查询 $Q = q_1q_2\cdots q_n$ 和文档 $D = d_1d_2\cdots d_m$，本文用 $p(d|q)$表示 Q 生成 D 的条件概率。由 Bayes 公式可得 $p(d|q)\propto p(q|d)p(d)$。如果将 $p(d)$视为统一值，则 $p(q|d)$的大小就是决定文档与查询相关度的依据。文本搜索一般采用 1 元模型(1-gram)，即假设单词之间是相互独立的，因此有 $p(q\mid d)=\prod_i p(q_i\mid d)$。其对数表示如下

$$\log p(q\mid d)=\sum_{i:c(q_i;d)>0}\log\frac{p_s(q_i\mid d)}{\alpha_d p(q_i\mid C)}+n\log\alpha_d+\sum_i\log p(q_i\mid C)$$

其中，$p_s(q_i|d)$表示在文档中出现查询词的频率，$p(q_i|C)$表示在文档集合中出现查询词的频率，α_d 是参数。为了防止条件概率为零，可引入平滑方法。文献[15]介绍了 3 种平滑方法，如表 2 所示。

表 2　3 种平滑方法的数学表示及其参数

平滑方法	$p_s(q_i\mid d)$	α_d	参数
线性插值平滑	$(1-\lambda)p_{ml}(w\mid d)+\lambda p(w\mid C)$	λ	λ
狄利克雷平滑	$\frac{c(w;d)+\mu p(w\mid C)}{\Sigma_w c(w;d)+\mu}$	$\frac{\mu}{\mid d\mid+\mu}$	μ
绝对折扣平滑	$\frac{\max(c(w;d)-\delta,0)}{\Sigma_w c(w;d)}+\frac{\delta\mid d\mid_\mu}{\mid d\mid}p(w\mid C)$	$\frac{\delta\mid d\mid_\mu}{\mid d\mid}$	δ

其中，$p_{ml}(w\mid d)=\frac{c(w;d)}{\Sigma_w c(w;d)}$。

文献[16]列出了信息检索模型的 7 条约束，在用词袋(bag of words)表示文档的基础上，从 TF、IDF、文档长度 3 个方面设定限制条件，在评价各模型满足限制条件的程度的同时改进模型以提升搜索效果。另外，融入语义信息的主题模型(topic model)[17]以及词嵌入(word embedding)[18]等也逐渐被应用到文本搜索中。另外，基于知识库的问答式搜索[19]与任务型搜索[20]也正成为研究热点。

3.3.2　短文本

短文本搜索在社交网络（如微博等）兴起后逐渐成为了一个重要的研究领域。如果长文本在企业中对应各类文件、产品说明书等，短文本则对应企业用户的微博、电商平台的用户评论等。短文本相比于长文本来讲，字数少但是信息量仍然大。因此，短文本的搜索更注重语义信息，并且短文本搜索大部分用于情感分析和意见挖掘。

短文本搜索也需要对文档进行表示，由于词频不是最重要的因素，所以可以借鉴意见挖掘中的相

关算法。以用户评论为例，文献[21]从语义角度对短文本进行表示，将其表示成(o_j, f_{jk}, oo_{ijkl}, h_i, t_l)，其中，o_j表示意见指向的对象，f_{jk}表示该对象的第k个特征，oo_{ijkl}表示意见内容，h_i表示表达意见的主体，t_l表示意见发表的时间。以此五元组表示短文本的特征，有助于意见分类，也有助于搜索。除此之外，POS(part of speech)标记等方法也可用于短文本表示。

3.3.3 多媒体搜索技术

全媒体时代，数据以文本、图片、视频、语音等格式在个人电脑和移动设备间产生和传播，文件搜索从单一模态搜索逐渐向多模态融合搜索过渡。其中，多模态融合主要集中在文本和图片的融合：一方面，深度学习在图片搜索领域不断突破，递归神经网络等初步实现了图片与文本的转化[22]；另一方面，基于新闻门户或者 Flickr、微博等社交平台的图文数据，可以实现比查询检索更高一级的事件检测[23]，通过图像特征和文本特征的相互补充，用半监督或无监督的方法将属于同一事件的内容聚类。企业拥有的数据，尤其是外部数据横跨各种模态，因此多模态信息检索也是企业级搜索面对的重大挑战之一。

4 智慧搜索

4.1 概念

一直以来，搜索算法与技术的发展一直处于几大类模型的框架内。即使是深度模型在自然语言处理中不断突破，其目的还是增强机器对文本的理解与表达，从而更有效地匹配查询与文档。文献[24]提出了“数据—信息—知识—智慧”的概念，如图 1 所示。其中，数据是对客观世界的记录，信息是增加了背景的数据，将信息提取规律后得到知识，而将（领域）知识应用到实际就是一种智慧。如果说信息已经解决了 who、what、where、when 的问题，知识能解决 how 的问题，那么智慧则是要解决 why 的问题。传统的搜索方式把和查询有关的文档按序罗列出来，从广度上实现了信息层的功能，但不能解决 how 和 why 的问题，即“知其然而不知其所以然”。而智慧搜索可以将搜索扩大到知识层和智慧层，让搜索能够返回比文档更直接的深度信息（知识及智慧）。

对搜索模型的研究意义固然重大，但对理解搜索意图的研究同样是优化搜索结果的重要内容。根据用户在搜索任务中的参与度，可将其分为用户引导式搜索、用户辅助式搜索和用户模糊式搜索。用户引导式搜索是指用户在查询任务中占主导作用，从查询分解、数据源选取到结果集成，用户都参与。其中，用户辅助式搜索是交互式的搜索，比如查询算法会在查询过程中与用户交互，及时更新领域知识和调整查询方向。用户模糊式搜索是智慧程度最高的搜索形式，即领域知识、搜索算法、数据源等都会在查询过程中自我学习、自我更新，增强查询的自动化和智能化。

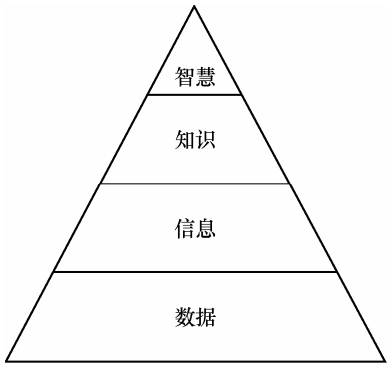

图 1 “数据—信息—知识—智慧”层次

大众在互联网上的主要搜索行为有查询、导航、交易等，但还有一大部分搜索是以提问题的形式，是一种系统性的搜索。搜索主体的意图不是简单的查询，更像是一种征询，如“机器学习如何入门？”、“与信息检索相关的技术有哪些？”等。如果用搜索引擎搜索这些词条，需要进行多次搜索才能得到比较满意的结果。当前的知乎、Quora 等网站是以众包的形式，集结网民的力量来解决这类搜索问题，但缺点是容易存在结果缺乏专业性、内容质量低下等问题。如何从大量的数据中提取符合搜索主体意图的系统化、高质量的信息也是目前搜索技术亟需要突破的一个方向。

企业中同样面临此类问题。比如对决策者而言，大多数的搜索实际上都是系统性的搜索。决策者的搜索意图往往不是单个数据源就能解决，而是需要整合多个数据源的数据进行分析与综合。员工资料、上一季度的产品销售额、下一季度的订单等，用传统的搜索方式都可以实现，但是某产品的市场反应如何、销量如何、销售为什么好、销售为什么不好等问题却无法依靠传统搜索方式解决，因为解决这些问题还需要综合市场、销售、生产等各部门的数据。理解搜索主体的意图就是为了能更准确地从大量的数据源中筛选出有用的数据，而这就需要引入

领域知识。

当前已有商业智能公司开发出一些智慧搜索产品。台湾的硕网资讯利用自然语言处理技术开发出智慧搜索、相关性引擎、语义分析、智能分类、文字勘探等核心技术，其中，WiSe 智能搜索引擎可以通过商业资料库、档案系统、企业邮件、文档图片、网页截取等非结构化资料，建立涵盖大数据的索引，为企业提供精确有效的搜索服务。

总而言之，智慧企业中的智能搜索要解决的不仅仅是搜索什么，更是如何去搜的问题。通过综合纵向搜索、横向搜索和“泛”搜索，让企业不仅能够快而准地搜索到企业内部和企业间的数据，还能够广而全地搜索到外部数据，使企业不仅能解决日常搜索的需求，还能增强对日常事务的“学习”，在企业的自我纠错、自我计划、自我发现、自我创造中发挥重要作用。

4.2　多维搜索法

4.2.1　“八爪鱼”的解释

如前所述，智搜不是一种单一的搜索算法，而是一个框架搜索的概念，涉及到对查询的扩展与细化。它从一个问题出发，启发式地将问题逐级分解到子问题，对于每个子问题，如果有领域知识或者第一手的数据源可利用，则直接用常规的方法从对应数据源中搜索相关的数据。如果没有，则需寻找更广阔的数据源，并采取特定的方式将返回的内容集成，得到最终结果。这样一个自顶向下分解任务和自底向上集成结果的过程构成智搜的主要步骤。对搜索意图的理解就是分解搜索任务，是将查询重新定义的过程，将其形象化称作多维“八爪鱼”搜索法。下面通过一个例子来解释它的机理。

以互联网金融行业中的个人征信为例，对于 P2P 网贷企业而言，用户的信用评级十分重要。如果要查询某一用户的信用级别，就需要对信用的概念进行扩展和细化。比如首先将个人信用评级依据其评判标准分解成身份验证、还款能力评估、还款意愿评估和好友信用 4 个维度（子问题），如图 2 所示。身份验证旨在确认申请贷款者的身份真实性及其提供的各项资料的可信度，需要从姓名、性别、职业、生日、地址、电话等特征验证，搜索的数据源是用户上传信息、证件信息、社交账户信息等；还款能力评估旨在评价用户负担还款的能力，需要从个人收入、账户的类型、储蓄账户余额、信用账户额度等特征进行建模，搜索的数据源是银行账户资料、网上支付平台账户资料等；还款意愿旨在根据用户的信誉历史量化用户的还款意向，需要从信用账户的数量、信用历史、贷款利率、还款记录、在线交易记录等特征衡量，搜索的数据源是信用卡账户资料、账单流水、网上购物记录等；好友信用旨在将用户的好友信用评价作为对用户本人信用预测的参照，需要收集好友亲近程度、好友信用状

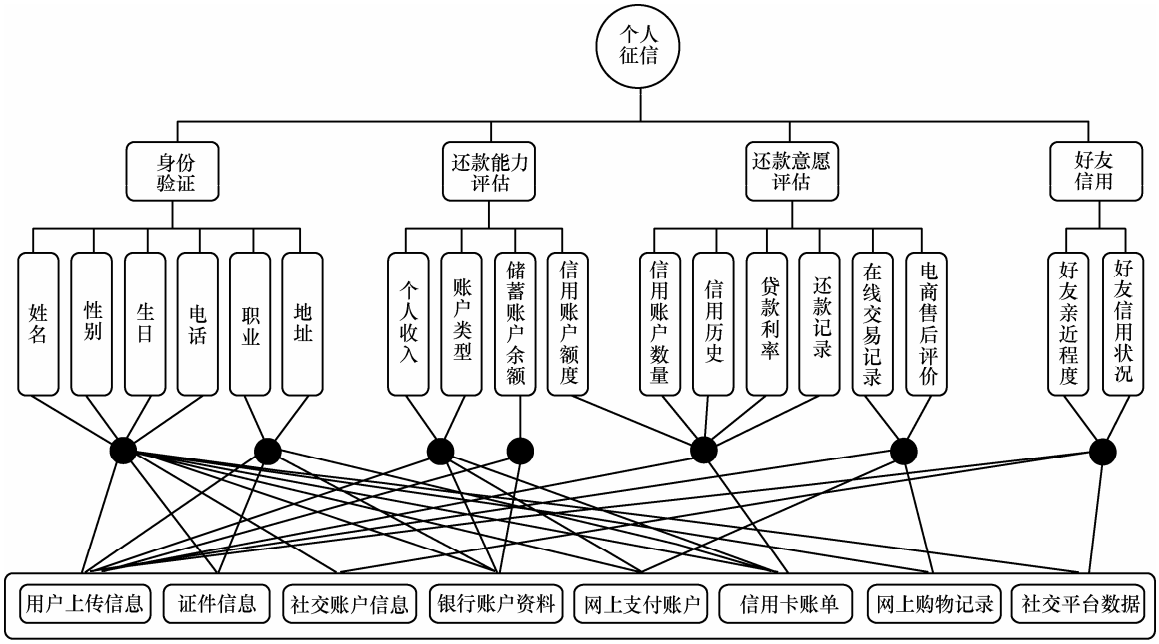

图 2　个人征信查询的多维“八爪鱼”搜索法

况等特征，搜索的数据源主要是社交网络。在个人征信中，搜索问题的重点是确定那些反映个人信用的特征维度并对其做分类，即 4.1 节中的搜索意图。确定了搜索的框架，就可以针对子问题的特点搜索相应的数据源。在这一搜索过程中，身份信息、银行账户、交易记录等数据是传统银行等金融机构贷款给个人时所参考的重要数据，这类数据的规律在互联网金融发展以前就为人所熟知，因此对这类数据的搜索实际上是将线下的工作转移到线上，特点是快速而精准；而网络支付平台、电商平台、社交网络的数据属于广泛而杂乱的数据，传统的经验不足以解决，因此需要借助“泛”搜索，实现多源异构数据的跨平台搜索。

4.2.2 多维搜索引出的挑战

多数据源的特点会给多维“八爪鱼”搜索法带来了一些新的挑战。第 1 个挑战是多源数据的格式差异。由于智慧企业中存在结构化、半结构化、非结构化数据，所以同一个查询可能对应不同类型的数据源，除了搜索方法不同，返回的数据格式也可能各异，因此集成结果之前首先需要实现数据的标准化。在个人征信的实例中，本文需要对所有搜索结果作标准化，特别是对信用历史、还款记录、在线交易记录、电商售后评价、好友亲近程度、好友信用状况等非结构化数据的量化。第 2 个挑战是多源数据的数据质量。数据质量管理是统计学研究的重要内容之一，数据的准确性、相关性、可靠性、时效性和完整性等是统计学家衡量数据质量的重要维度。以社会调查为例，无论是传统的问卷调查还是现代的网络调查，由于受访者背景不一、调查区域的局限、问题设置的科学性存疑等原因造成的回收数据质量参差不齐。在多维“八爪鱼”搜索法中，多源数据同样存在数据质量问题。下面分别从多源数据质量评估和缺失数据跨源补充两方面做详细说明。

多源数据质量评估是从搜索结果的准确性、可靠性和相关性 3 个维度来评估数据质量：1) 针对标准化后的结果准确性，验证需要对内满足领域知识预先设定的数据格式、范围等，对外比较不同数据源的结果以保证其准确性；2) 针对数据的可靠性，评价可通过为每个数据源先设定可靠因子，比如初始默认每个数据源的可靠因子相等，一旦有错误、虚假或不合理事件发生，就要降低对应数据源可靠因子的值；3) 对结果进行相关性检验则是建立在查询间相互关联的基础之上，数值型查询可通过历史数据学习线性或非线性模型，将模型的预测结果与搜索结果比较，以判断搜索结果是否合理，非数值型查询的相关性检验需要结合领域知识。

缺失数据补充并非“智搜”里面才有的问题，但在“智搜”场景下具有其独特性。对一些难以再分解但仍缺少数据源或数据源内容不足的查询，需要解决数据的完整性问题，而不同类型的缺失数据其补充方法也会有所不同，包括：1)直接根据领域知识预测缺失数据；2)深入本数据源或其他数据源搜索并查询有关的隐含结果；3)根据该查询与其他查询的关联性，使用回归或分类模型来预测该查询的结果；4)从查询历史中找到相似的实例并用该实例中对应的数据作为间接补充。

4.2.3 多维“八爪鱼”搜索法

下面给出多维“八爪鱼”搜索法的一般化描述。

Q_0：初始查询（默认非空）；

$Q_{0l_1\cdots l_{m-1}l_m}$：查询 $Q_{0l_1\cdots l_{m-1}}$ 的第 l_m 个子查询，l_m=1,⋯,N_m, N_m 是 $Q_{0l_1\cdots l_{m-1}}$ 的子查询数量；

Q_{leaf}=$\{q_1,\cdots,q_L\}$：查询树节点的集合；

W=$\{w_i\}$, $w_i=\{w_{iS_{n_1}},\cdots,w_{is_{n_i}}\}$, Q_{leaf} 对应数据源的可靠因子集合,i=1,2,⋯,L, n_i 表示与 q_i 对应的数据源总量；

T：历史查询实例的数量；

$F_{is_{n_h}}$：历史查询实例中，查询 q_i 在数据源 S_{n_h} 的结果不可靠的实例数量；

R_{leaf}=$\{r_i\}$, $r_i=\{r_{iS_{n_1}},\cdots,r_{iS_{n_i}}\}$：查询树叶节点返回搜索结果的集合；

$R_{0l_1\cdots l_{m-1}l_m}$：$R_{0l_1\cdots l_{m-1}l_m}$ 的查询结果；

R_0：查询最终结果；

S=$\{S_1,\cdots,S_K\}$：所有数据源；

D：领域知识。

1) 自顶向下分解查询

①如果 S 中有与查询 Q_0 直接相关的数据源 $\{\cdots,S_j,\cdots\}$或查询 Q_0 不可再分，则该节点无需再做分解。

②如果 S 中没有与查询 Q_0 直接相关的数据源，则引入领域知识 D，将查询 Q_0 分解为若干个子查询 $\{Q_{01},Q_{02},\cdots,Q_{0N_1}\}$。

③如果 S 中有与查询 $Q_{0l_1\cdots l_{m-1}}$ 直接相关的数据源 $\{\cdots,S_j,\cdots\}$或该查询不可再分，则该节点无需再做分解。

④如果 S 中没有与查询 $Q_{0l_1\cdots l_{m-1}}$ 直接相关的数据源，则引入领域知识 D，将查询 $Q_{0l_1\cdots l_{m-1}}$ 分解为若干个子查询 $\{Q_{0l_1\cdots l_{m-1}},\cdots,Q_{0l_1\cdots l_{m-1}N_m}\}$。

⑤递归调用③～④。

⑥得到查询的树状结构。

2) 自底向上集成结果

①如果查询 Q_0 的树状图只有一个节点，进入下一步，否则进入③。

②如果是单一数据源，则返回查询最终结果 R_0；如果有多个数据源，则根据查询本身的性质，执行多源数据质量评估和缺失数据间接补充等，返回查询最终结果 R_0。

③查询结束。

④对叶节点查询 $q_i \in Q_{\text{leaf}}$，重复②，得到对应结果或列表。

⑤自底向上回溯，对非叶节点查询 $R_{0l_1\cdots l_{m-1}l_m}$，假设其子节点的查询结果是 $R_{0l_1\cdots l_m1},\cdots,R_{0l_1\cdots l_mN_{m+1}}$，则 $R_{0l_1\cdots l_m}=assemble(R_{0l_1\cdots l_m1},\cdots,R_{0l_1\cdots l_mN_{m+1}})$，$assemble$ 函数是对下一层子查询结果的直接集成。得到该节点的查询结果 $R_{0l_1\cdots l_{m-1}l_m}$。

递归调用⑤，直到抵达 Q_0，返回查询最终结果 R_0，查询结束。

3) 多源数据质量评估

①初始化 w_i 为全 1 集合。

②查询正式运行后，有新的查询，T 加 1。

③数据准确性验证——验证同一个查询从不同数据源中返回的搜索结果是否相同。如果 q_i 在数据源 S_{n_h} 中的搜索结果确认不准确，则 $F_{iS_{n_h}}$ 加 1。

④数据可靠性评价——利用查询历史数据量化数据源对叶节点查询的可靠性。数据源 S_{n_h} 对查询 q_i 的可靠性可量化为 $w_{iS_{n_h}}=\dfrac{T-F_{iS_{n_h}}}{F_{iS_{n_h}}}$。数值总和型的搜索结果 $r_i=\sum_{h=1}^{i}w_{iS_{n_h}}r_{iS_h}$，数值期望型的搜索结果 $r_i=\dfrac{\sum_{h=1}^{i}w_{iS_{n_h}}r_{iS_h}}{\sum_{h=1}^{i}w_{iS_{n_h}}}$。

⑤数据相关性检验——通过检验相同深度但不同维度的查询之间的相关性判断搜索结果是否合理。对于数值型的查询结果 $R_{0l_1\cdots l_{m-1}i_1}$，如果它与同一深度的其他结果 $\{R_{0l_1\cdots l_{m-1}i_2},R_{0l_1\cdots l_{m-1}i_3},\cdots,R_{0l_1\cdots l_{m-1}i_k}\}$ 存在线性或非线性关系，那么可以从历史数据训练得关系模型，将搜索结果与用关系模型得到的数据比较，判断结果是否合理。对于非数值型数据，则需要领域知识。

在上述的个人征信例子中，姓名、性别、生日、电话、职业、地址等都需要确保各数据源的搜索结果相同，否则说明存在错误或虚假信息。个人收入、储蓄账户余额、信用账户数量等使用数值总和型搜索结果，信用账户额度、贷款利率等可以使用数值期望型搜索结果。由于个人收入与个人背景、储蓄余额及信用额度之间存在一定的关联性，因此可以借助历史数据，在给定个人背景、储蓄余额及信用额度下用回归模型来预测个人收入，并与实际搜索结果做比较。

4) 缺失数据间接补充

缺失数据的补充分多种情况。

①直接根据领域知识预测缺失数据。

②再次深入本数据源或其他数据源搜索隐含的与查询有关的结果。

③根据该查询与其他查询的关联性，使用回归或分类模型预测该查询的结果。

④从查询历史中找到相似的实例对象并用该实例中对应的数据作为补充。

比如在个人征信应用中，如果查询结果有缺失，可以利用领域知识、数据源、不同维度和相似对象等来作间接补充，包括从网购记录来推测用户性别，用个人背景、储蓄余额和信用额度来推测个人收入，还可以凭相似用户的职业来推测查询用户的职业等。

4.3 进一步的工作

智慧搜索的数据范围广泛，涵盖了产品的研发、生产、销售等各个环节，囊括了订单、产品说明书、社交网络等各种类型的数据。搜索方式除了 4.2 节中提到的多维“八爪鱼”搜索法，还可以使用“圈子”搜索法，即划定一个以产品或者用户为中心的圈子；以产品为中心的圈子覆盖产品相关的人、事、物，以用户为中心的圈子则涉及用户与企业、用户与产品、用户与用户之间的关系。因此搜索脉络可以建立在圈子内的基础上，也可以建立在圈子间的基础上，查询目标不同，搜索脉络也不尽相同。

5 结束语

信息技术的发展使人类有能力记录、收集、创

造和存储，同时分享和传播更广泛的数据，降低了大众获取信息的门槛，也节约了信息流通的成本，然而，大数据带来的并非全都是优势，比如低质、杂乱、大量的数据有时候反而会分散决策者的精力，因此，在不破坏数据完整性的前提下对原始数据统一调配与规范化，将对提高决策速度和质量产生积极的效果。智慧企业的概念在企业信息化程度日益增强的今天走进公共视野。随着数据源不断增加、数据量快速上升、数据形式逐渐多元、企业对生产管理和决策智能化的要求不断增强，以数据为中心的搜索模式是在此背景下的必然选择。传统数据搜索的方法虽然高效但是单一，不足以解决知识层和智慧层的搜索任务。在本文中提出的多维“八爪鱼”搜索，目的是在传统搜索的基础上展开一种新的思路，旨在为智慧企业的实现与发展增加一种新的知识服务。通过将查询由关键词扩展到问题或任务，将返回的结果由文档扩展到方案或步骤，笔者期望智慧搜索在企业级应用中成为企业制定决策不可或缺的工具，从而切实帮助企业更加科学、有效地解决问题，实现真正意义上的智慧企业。

参考文献：

[1] MANNING C D, RAGHAVAN P, SCHÜTZE H. Introduction to Information Retrieval[M]. Cambridge: Cambridge University Press, 2008.

[2] GUO L, SHAO F, BOTEV C, et al. XRANK: ranked keyword search over XML documents[A]. ACM SIGMOD’03[C]. San Diego, CA, 2003.9-12.

[3] XU Y, PAPAKONSTANTINOU Y. Efficient keyword search for smallest LCAs in XML databases[A]. Proceedings of ACM SIGMOD’05[C]. Baltimore, Maryland, USA, 2005.14-16.

[4] LI G, FENG J, WANG J, et al. Effective keyword search for valuable lcas over xml documents[A]. Proceedings of ACM CIKM’07[C]. Lisboa, Portugal, 2007. 6-8.

[5] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[6] TURNEY P D, PANTEL P. From frequency to meaning: vector space models of semantics[J]. Journal of Artificial Intelligence Research, 2010, 37(1): 141-188.

[7] FUHR N. Probabilistic models in information retrieval[J]. The Computer Journal, 1992, 35(3): 243-255.

[8] SONTAG D, COLLINS-THOMPSON K, BENNETT P N, et al. Probabilistic models for personalizing web search[A]. Proceedings of the fifth ACM WSDM conference[C]. Seattle, Washington, 2012. 8-12.

[9] PONTE J M, CROFT W B. A language modeling approach to information retrieval[A]. Proceedings of ACM SIGIR’98[C]. Melbourne, Australia, 1998.275-281.

[10] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011, 12(1): 2493-2537.

[11] WU H C, LUK R W P, WONG K F, et al. Interpreting TF-IDF term weights as making relevance decisions[J]. ACM Transactions on Information Systems (TOIS), 2008, 26(3): 55-59.

[12] SINGHAL A, BUCKLEY C, MITRA M. Pivoted document length normalization[A]. Proceedings of ACM SIGIR’96[C]. Zurich, Switzerland, 1996.21-29.

[13] ROBERTSON S, ZARAGOZA H. The probabilistic relevance framework: BM25 and beyond[J]. Foundations and Trends in Information Retrieval, 2009, 3(4): 333-389.

[14] DING G, BAI S, WANG B. A survey of statistical language modeling for text retrieval[J]. Journal of Computer Research and Development, 2015, 43(5):769-776.

[15] ZHAI C, LAFFERTY J. A study of smoothing methods for language models applied to information retrieval[J]. ACM Transactions on Information Systems (TOIS), 2004, 22(2): 179-214.

[16] FANG H, TAO T, ZHAI C. Diagnostic evaluation of information retrieval models[J]. ACM Transactions on Information Systems (TOIS), 2011, 29(2): 217-230.

[17] BLEI D M. Probabilistic topic models[J]. Communications of the ACM, 2012, 55(4): 77-84.

[18] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv Preprint arXiv: 1301.3781, 2013.

[19] YIH W, CHANG M W, HE X, et al. Semantic parsing via staged query graph generation: question answering with knowledge base[A]. Proceedings of 53rd Annual Meeting of ACL and the 7th IJCNLP[C]. Beijing, China, 2015.

[20] YANG Z, NYBERG E. Leveraging procedural knowledge for task-oriented search[A]. Proceedings of ACM SIGIR’15[C]. Santiago, Chile, 2015.

[21] LIU B. Sentiment analysis and subjectivity[J]. Handbook of Natural Language Processing, 2010, 30(36):152-153.

[22] ANDREJ K, LI F. Deep visual-semantic alignments for generating image descriptions[J]. arXiv Preprint arXiv:1412.2306, 2014.

[23] YANG Z, LI Q, LU Z, et al. Semi-supervised multimodal fusion model for social event detection on Web image collections[J]. International Journal of Multimedia Data Engineering and Management (IJMDEM), 2015, 6(4): 1-22.

[24] ACKOFF R L. From data to wisdom[J]. Journal of applied systems analysis, 2010, 16: 3-9.

作者简介：

陈扬斌（1992-），男，浙江金华人，香港城市大学硕士生，主要研究方向为机器学习和数据挖掘。

李青（1962-），男，浙江金华人，香港城市大学教授、博士生导师，主要研究方向为数据仓库、多媒体检索、网络挖掘等。

庄越挺（1965-），男，浙江慈溪人，浙江大学教授、博士生导师，主要研究方向为多媒体信息检索、人工智能、大数据技术等。