电子信息类新技术丛书

软件人主题分析与信息检索技术

周亦鹏◎著

北京邮电大学出版社
www.buptpress.com

电子信息类新技术丛书

软件人主题分析与信息检索技术

周亦鹏 著

北京邮电大学出版社
www. buptpress. com

内 容 简 介

随着移动应用、社会网络应用的快速发展，用户随时随地获取个性化信息的需求更加强烈，对检索系统的智能化要求更高。软件人以其拟人化的特征更好地适应了这一发展趋势，同时通过与物联网相结合，软件人的感知能力大大提高，这使软件人成为实现智能化、拟人化主题分析和信息检索的一种重要途径。

本书首先分析了主题分析和信息检索技术的现状，进而探讨了软件人在互联网信息处理发展演化中所能起到的作用。重点阐述了软件人的构造，尤其是以主题模型来建立软件人的语言模型，使之能够模拟不同人的语言模式。围绕软件人的主题模型，从文本主题分析、主题模型的语义标注、跨媒体主题分析几个方面总结了作者在该领域的理论研究工作。在理论研究的基础上，进一步介绍了软件人主题分析技术在食品安全事件监测、智慧旅游和领域主题信息检索中的应用。

本书可作为从事智能科学技术、计算机、信息检索相关的科研、教学和工程技术人员参考用书，也可作为高等院校的专业用书。

图书在版编目(CIP)数据

软件人主题分析与信息检索技术/周亦鹏著. --北京：北京邮电大学出版社，2012.8
ISBN 978-7-5635-3218-6

Ⅰ. ①软… Ⅱ. ①周… Ⅲ. ①主题分析②情报检索 Ⅳ. ①G254.21②G252.7

中国版本图书馆 CIP 数据核字(2012)第 205241 号

书　　名：软件人主题分析与信息检索技术
作　　者：周亦鹏
责任编辑：何芯逸
出版发行：北京邮电大学出版社
社　　址：北京市海淀区西土城路 10 号(邮编：100876)
发 行 部：电话：010-62282185　传真：010-62283578
E-mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：北京联兴华印刷厂
开　　本：720 mm×1 000 mm　1/16
印　　张：12
字　　数：234 千字
版　　次：2012 年 8 月第 1 版　2012 年 8 月第 1 次印刷

ISBN 978-7-5635-3218-6　　定　价：29.00 元

前　言

随着网络与计算机技术的迅速发展，Web 信息爆炸性地增长，互联网已经成为一个巨大的海量信息空间。根据 ComScore 公司的统计，2008 年 12 月底，全球互联网用户总数突破 10 亿，其中 41.3%的用户来自亚太地区。中国互联网用户所占的比例为世界第一，达到 17.8%。截至 2008 年年底，中国的网站数达到 287.8 万个，网页总数超过 160 亿个。Google 在 2008 年索引的 Web 网页规模已经超过了 1 000 亿，海量的 Web 信息极大地丰富了人们的生活，但也给信息的组织、查找与分析带来了极大的挑战。

如何快速、准确、方便地从海量 Web 信息库中获取所需要的信息，一直是人们关心的一个重要问题。为此，近十年来研究人员和业界做出了极大努力，取得了巨大成果，发明了如 PageRank，HIT 等著名算法和竞价排名等商业模式，涌现出 Google、百度等家喻户晓的搜索引擎公司，不仅在相当程度上解决了 Web 信息获取的问题，也为计算机科学与技术的发展拨开了一片新的空间。

尽管如此，随着 Web 的演化（Web 1.0→Web 2.0→Web 3.0），不仅其信息量继续在以指数率增长，而且信息产生的方式、种类和风格也趋于多样，向 Web 搜索技术不断提出新的挑战。未来的搜索引擎将不再只是按词语的出现对网页进行简单索引，而将是基于对 Web 信息集合进行深度分析和挖掘后的知识展现。目前 Web 数据的搜索与挖掘主要是基于 Web 文档与链接结构处理，对于 Web 数据中复杂信息模式与信息关联的表示存在着明显的不足。具体表现在语义表达多样性；数据组织缺乏关联信息；单一媒体形式的检索，缺乏融合视觉、听觉、语言处理技术的有效跨媒体检索手段。

在目前从 Web 2.0 到 Web 3.0 的发展过程中，可以看到 Web 数据量越来越大，信息种类多样化及其关联复杂化。因此上述存在的问题变

得越来越突出，迫切需要探索 Web 搜索的新理论与方法，设计新的概念、模型、机制和算法，使搜索引擎能够更快、更准、更方便地从更加复杂的海量 Web 信息库中获取所需要的信息。

目前，搜索引擎对人的活动，尤其是人的社会化活动的支持还很有限。社会网络应用（如 MySpace，Facebook，Flickr 和 YouTube）最近几年在这方面有着极大的发展。这些应用主要在 Web 平台上构建了人与人交流的社区，协助具有类似兴趣爱好的人之间进行交流，方式包括即时消息、E-mail、聊天室、讨论者、文件共享等。这些应用成功的关键在于它们在网络空间内提供了一个人与人交流的平台。同时，这些社会网络应用的快速发展也反映了现代社会中人与人之间存在迫切交流的需要。因此，对于以搜索引擎为代表的信息检索系统来说，其发展步伐已经落后于社会网络应用，要真正实现“以人为本”的目的，还应该更多的从人的角度来看待系统的要求。

在“以人为本”的研究中，“软件人”是更加人性化的软件实体，它是拟人的软件人工生命，是在特性、功能、行为、结构、组织等方面对人及人类社会的模拟、延伸和扩展。软件人最初定义为在网络环境下的虚拟机器人，它融合了人工智能、人工生命、分布式计算、智能体和计算机网络等领域的理论和技术。但是，随着物联网的发展，现实物理环境呈现出与虚拟数字环境的融合趋势，这一融合的渠道就是互联网。因此，作为存在于互联网环境中的软件人，借助于物联网的支持具有了更强的感知现实世界，甚至改造物理环境的能力。软件人技术的发展为实现以人为中心，服务于人的智能化信息检索提供了可能。这种信息检索是通过感知人的活动情境和意图，并且通过拟人化的社会活动来主动提供个性化的信息服务，将大大改善信息检索的服务体验。

本书给出了信息检索及其关键技术——主题分析的发展趋势，以及软件人在互联网信息处理发展演化过程中可能形成的社会网络结构，详细论述了软件人主题分析及信息检索理论，介绍了软件人主题分析和信息检索在食品安全事件监测、智慧旅游和领域主题信息检索中的应用。

本书内容分两部分，共 9 章。第一部分为软件人主题分析和信息检索理论，包括第 1 章互联网主题分析和智能信息检索，第 2 章软件人和互联网信息处理，第 3 章互联网环境下的软件人及其语言模型，第 4 章

文本情境主题分析，第5章情境主题模型的自动标注方法，第6章跨媒体主题分析方法；第二部分为软件人主题分析的应用，包括第7章主题分析在食品安全事件监测中的应用，第8章软件人主题分析在智慧旅游中的应用，第9章软件人在领域主题信息检索中的应用。

本书在资料整理和写作过程中得到了杜军平教授的大力支持，以及胡娟老师，梁美玉、韩鹏程、紫玲玲、杨月华等博士的帮助，在此向他们表示感谢。

本书的写作和出版得到了北京工商大学学术专著出版项目的支持和资助，特此致谢。

由于作者水平有限，加之时间仓促，书中不妥之处在所难免，恳请读者批评指正。

作 者

2012年6月于北京工商大学

目　　录

第1章　互联网主题分析和智能信息检索

1.1　文本主题分析

目前，互联网上无论是网页、博客、邮件、即时信息还是微博，仍然是以文本信息为主，因此文本挖掘仍然是主题分析的重要手段。

(1) 主题建模与主题识别

主题分析最早来自于对文本主题结构和主题表达的研究，主要方法是利用词语出现的频率和分布等统计信息计算词语及句子的相对重要程度，提取并输出能够反映文本中心思想的信息，从而获得文本主题。其中，识别文本主题是主题分析的重要研究内容，主要通过关键主题词集合、关键短语集合和关键句子集合等方式表达文本主题。目前的主题识别方法主要分为两类：一类是基于统计的方法，另一类是基于语义分析和理解的方法。基于统计的主题识别方法利用词语在文本中出现的频率、分布和共现信息等统计信息找出最能表达文本主题的多个词语组成主题词集合。基于语义分析和理解的方法则是利用知识库和人工智能技术对文本进行深层的句法和语义分析，产生连贯的符合文体要求的主题。由于语义分析需要庞大的知识库和完善的语言学规则，此类方法只能应用于特定题材的文体和内容，使用受到极大限制。而基于统计的主题识别方法借鉴信息检索中的 TF-IDF、Bayes 等方法计算词语的权重并抽取关键词，可以避开语义分析的难点，因此一直是最常用的方法。

利用统计方法进行主题分析必须选择合适的语言模型。有限混合模型可以代表文本中的词汇分布，并直接利用 EM 算法进行训练，但会出现局部极大值和收敛速度过慢的问题，且错误率较高。PLSA 是另一可选模型，但模型中的文档概率值与特定文档相关，因此缺乏处理新文档的自然方法，且模型易于出现过度拟合的问题。与 PLSA 模型相比，LDA 是完全的生成模型，它将主题混合权重视为 k 维参数的潜在随机变量，而非与训练数据直接联系的个体参数集合，推理上采用拉普拉斯近似、变分近似以及期望—扩散等方法获取待估参数值，克服了上述不足。

但是，传统的主题分析方法假定整个数据集只具有一个统一的语言模型，忽略了情境变化对语言模型的影响。研究人员对 Web 2.0 应用中用户发布的文本信息进行了研究，发现不同用户、不同应用平台发布的内容在语言模型、词语分布等方面表现出了多样的特征。这些内容往往依赖特定的语言环境，比较典型的是微博由于字数限制，属于一种受限的语言环境，单从信息内容本身进行挖掘难以全面准确地获得有用的知识。此时必须利用各种情境数据，如时间、空间、事件背景、用户特征及应用平台等，来发现不同情境下的特征并建立相应的语言模型。

因此，好的主题分析模型应该能够混合多样化的情境因素。虽然混合概率模型可以将每个主题分别基于相应的概率模型进行特征描述，但是这种方法缺乏对主题以外的其他情境信息建模的能力，且无法将先验知识引入主题模型，无法确定其他情境对主题本身的影响。最近一部分研究已经开始尝试将其他情境信息，如将作者、时间等集成到主题模型中。

（2）主题发现与跟踪

主题分析技术在主题发现与跟踪领域得到了大量应用，成为互联网舆情分析的重要基础。主题发现与跟踪是以大规模的文本语料为研究对象，通过分析文本内容所描述的主题，来发现新的主题内容，或者对属于原有主题的新的相关内容进行跟踪，最后将涉及某个主题的内容组织起来并且以一定的方式进行展现。与传统的信息检索技术相比，主题发现与跟踪技术更关注对新主题（如新事件和新话题）的发现，并且更强调获得新信息与特定主题的相关性。主题发现系统分为全自动方式和半自动方式两种，代表性工作包括 Rutgers 大学的 HDDI、马萨诸塞大学的 TimeMines 等全自动系统，以及 CIMEL 和 ThemeRiver 等半自动系统。

早期的主题发现研究主要集中在聚类方法的选择与融合上，考虑到主题表达的多粒度问题，利用层次化结构对主题进行发现与组织的方式也被提出来。随着互联网的应用和信息检索技术的发展，有研究将信息检索和机器学习技术相结合，用于主题发现与追踪，并在跨语言主题发现与追踪上取得了较好的结果，同时，网络论坛、博客上的主题挖掘开始出现。社会网络服务的发展使主题分析的研究开始将社会关系引入进来，Victor Cheng 等人提出了一种基于用户关系的主题挖掘模型，Mingliang Zhu 等人将词的相关性与用户相关性结合起来挖掘网络论坛上的话题。但是，主题分析领域面临的困难主要是如何建立主题信息的特征表示和相关性度量模型，特别是基于语义方法和考虑时空等情境变化的分析手段。

微博、博客等多种 Web 2.0 应用平台的出现，以及在这些平台上大量涌现、传播和相互关联的信息，使得主题信息的发展态势和传播规律成为关注的重点。主题信息的宏观态势分析关注关键信息点的量化分析与评价、主题信息在时间维度与空间维度上的发展态势分析与预测等。内容包括文本主题的涌现态势发现

(Emerging Trend Detection,ETD)和时序挖掘(Temporal Text Mining,TTM)。ETD 主要发现在一个时间范围内逐渐引起人们兴趣,并被越来越多的人使用、表达和传播的事件、概念等。

主题发现与跟踪领域面临的困难是如何建立跨平台跨媒体主题信息的特征表示和相关性度量模型,特别是要解决不同主题空间上主题模型间的相互转换和统一描述问题,利用跨平台跨媒体信息在高层语义上的关联关系实现主题发现与跟踪。

研究人员对 Web 2.0 应用中用户发布的文本信息内容进行了研究,发现不同用户发布的信息在语言模型、词语分布等方面表现出了多样的特征。这些内容往往依赖特定的情境,单从信息内容本身进行挖掘难以全面、准确地获得有用的知识。此时必须利用各种情境数据,如时间、空间、用户及其社会关系等,甚至通过进一步推理来发现不同情境下的特征并建立相应的语言模型。

在自然语言理解的研究中,上下文作为一种情境被用于监督学习,来扩展特征空间,进行实体抽取、标记等,也通过局部上下文的比较来进行语义消歧或实现词语聚类等。但主要问题是缺乏对更具一般性的情境挖掘方法的研究。常见的文本挖掘采用文本生成模型的方法,其核心思想是建立概率模型来描述观察到的文本数据的生成过程。这种文本数据的概率模型称为语言模型,常常被用来挖掘文本中的模式(如主题模式),以便于实现文本检索、分类、聚类或文本摘要等任务。传统的文本挖掘方法假定整个文本集只具有一个统一的语言模型,因此忽略了情境变化对语言模型的影响。有研究注意到这一问题,发现文本的统计信息(如词频)对于不同风格类型和体裁的文档和文档之间是有区别的。进一步研究发现好的文本集应能够混合多样化的情境因素,因此认为混合泊松分布比标准的泊松分布模型更适合文本数据,但是这些研究仅停留在文本集合级的建模,不能更为深入地理解情境的内容以及对特定情境进行建模。如何定义文本中的情境并将其特征进行描述,如何获得情境所产生的影响,如何利用情境信息解决各种主题分析任务仍然是留待解决的关键问题。

混合情境建模的最新进展是概率主题模型的研究,它的基本假设是文本数据通常覆盖了多个主题,因此可以通过混合概率模型建模,每个概率模型对应一个主题,同时一个主题也可以基于相应的概率模型进行特征描述。但是,这种方法缺乏对主题以外的其他情境信息建模的能力,且无法将先验知识引入主题模型,无法确定其他情境对主题本身的影响。最近一部分研究已经开始尝试将其他情境信息,如将作者、时间等集成到主题模型中。这其中主要的问题是缺乏一个一般性的框架,来定义情境并且描述建模的一般方法,无法根据情境对数据内容进行比较并从中发现情境模式。

1.2 视觉主题分析

随着多媒体数据逐渐成为网络信息的主要载体，尤其是各类热点主题事件进程中网络上大量出现的图像信息、视频信息，使基于文本的主题分析方法越来越难以全面、准确地挖掘相关主题内容及其影响范围和深度。研究融合各类媒体信息的跨媒体主题分析方法已经成为亟待解决的关键科学问题。

多媒体信息非结构化的数据格式、巨大的数据量以及内容理解的主观性与多义性，给多媒体数据的分析与理解带来很大的困难。近些年来，基于内容的多媒体分析、自动标注、索引构建和相似检索等研究一般是基于底层的视觉特征（如颜色、纹理、形状、运动、镜头等）和听觉特征（频谱分布和变化规律、节奏、韵律、话音、说话人特征等）。但这些底层特征并不能直接表示突发事件所蕴含的语义概念，也就是说在底层特征与高层语义之间存在语义鸿沟，关键是要将这些底层特征与高层语义关联起来。

虽然在图像的研究中也有一些基于视觉特征的聚类工作，但是由于难以抽出图像的基本语义单元，所以无法对图像进行语义分类。目前有两种基本对策，第一种对策是利用与图像相关的文本信息，将图像检索/分类问题转化为文本检索、分类问题，其缺点是完全舍弃了图像自身所包含的丰富信息，同时会受到文本处理能力的制约。而且对于图像和视频等形式的热点主题信息，一般难以用自然语言进行准确的描述，无法表达其实质内容和语义关系，所以这种挖掘方法精度比较低。第二种对策即图像自动语义标注，试图借助机器学习的方法，根据图像自身的视觉特征，建立关键词与可视语义对象之间概率意义上的某种关联以缩小语义鸿沟，这在本质上是一个大规模多分类问题。在过去几年中，相关研究获得了广泛关注，并取得了一定成绩。这些研究主要采用了同现模型、层次混合聚类模型、生成关联模型、翻译模型、潜层空间模型、SVM 分类器等方法。由于关键词集合可以代表一组主题语义概念，因此这项工作也可以视为对图像进行主题语义分类，但其性能会受到限制。

针对跨媒体数据挖掘领域的多模态特征融合等问题，已经有学者提出了一系列解决方法。文献[55]采用谱聚类方法分别建立了视觉表示空间和文本表示空间的二部图，为了保证聚类结果的一致性，将两个单独的聚类问题形式化为一个约束多目标优化问题。可以分别定义同一类型数据和不同类型数据之间的关系，将不同空间的相似度相互传播直到收敛，然后利用谱聚类算法在更新了相似性度量的视觉空间中进行聚类。有研究采用 Boosting 算法对图像分类进行了研究，建立了自然图像分类系统，并且通过提取人文、山川、植物、水域四维特征并结合相关数

据，实现了跨媒体信息导航。

近些年发展起来的翻译学习技术，实现了不同模态异构数据间的特征映射，如采用文本训练数据解决图像数据的认知学习问题，为解决跨媒体数据挖掘的语义鸿沟问题提供了一种新的途径。但是作为一种新兴的学习策略，目前翻译学习仅在较小的标注规模上有显著效果，因此需要设计一种高效的翻译学习算法以尽可能降低算法的计算复杂性。同时为了实现跨媒体挖掘，算法能力也有待进一步提高，以便将来自其他特征空间的有用特征映射到目标特征空间，改进目标特征空间的学习效果。

1.3　社会网络中的主题分析

目前社会网络服务(Social Network Service，SNS)已经成为 Web 2.0 的基础应用之一。SNS 网站及应用在全球蓬勃发展，在美国，Myspace，Facebook，Twitter 等发展迅速，中国本土的 SNS 网站也遍地开花，校内网、开心网等都拥有着庞大的用户群。据美国互联网调研公司 ComScore 研究报告显示，目前全球社交网站用户数量已增至 5.8 亿，与去年同期的 4.64 亿相比，增长 25%。根据 CNNIC 统计，截至 2009 年年底中国使用交友和社交网站的网民数达到了 1.76 亿。可见 SNS 已经是目前互联网行业最受瞩目的新兴领域。

与传统互联网服务不同的是，SNS 为人们提供更为社会化的应用和服务，从而使用户既可以发布自己的个人信息，也可以获取其他用户的信息，还可以与其他用户建立各种社会关系，形成 Web 社团，成员间可以进行多种形式的信息交换和共享。SNS 的出现改变了以往简单的依赖于文档链接的 Web 内容组织形式，来自不同内容发布者的不同主题的内容因为人的关注、评论、分享、标注等社会化行为被重新编辑、组织和传播。正所谓“物以类聚，人以群分”，因此 Web 内容的组织极大地受到了人与人之间的社会关系和社会活动的影响，同一网络社团中的人群所关注的主题会呈现出某种共性特征。在这种情况下，主题分析就不得不考虑网络中的社团结构及社团成员间的社会化行为。

目前，很多社会网络关系的研究是借助名词共现分析技术从 Web 数据中挖掘实体间的关系。例如，可视化在线社会网络的软件系统 Flink，通过分析 Web 页面、电子邮件信息、FOAF 文件来挖掘社会网络；Faloutsos 从 5 亿 Web 页面中，挖掘出一个 1 500 万人的社会网络；Knee 利用人名和音乐关键字在前 50 个 Web 页面并发的技术，把音乐艺术家分成了不同的等级。

此外，利用 Web 的链接结构和 Web 图模型来发现网络社区资源已经取得了一些成就。与此同时，也出现了各种以链接分析技术为基础的 Web 社区挖掘方

法，即将 Web 看作由页面和超链接构成的 Web 图，一个 Web 社区是一个 Web 图的子图，发现社区的过程就是从该 Web 图中找到一个适当的割集。通过用 HITS 方法分析链接结构可以获得 Web 社区，认为社区是由稠密的权威页面构成的核，把 Web 社区看作一些二分有向图的核。通过最大流算法也可以发现 Web 社区，认为社区是具有社区内页面之间的链接数大于同社区外页面之间的链接数这一特性的页面形成的集合。

目前，网络社区挖掘已经从社区中人和人的关系扩展到主题和内容的分析、个人爱好和信息的分析、组成结构分析等。例如，通过对群体关联性和兴趣分享的研究，实现个人行为的分析；基于最优线性组合的学习方法对多关联复杂社会网络中的隐藏信息进行挖掘；构建基于语义 Web 技术的社会网络抽取和分析系统，以及统一的社会网络语义模型 Ontotogies。但是，现有 SNS 网络社区的分析方法主要是在静态模型上进行挖掘，缺乏对实体间交互时间的考虑，无法动态反映网络社区的形成和用户兴趣的变化。

网络社区的存在使主题网页在 Web 上的分布服从一定的规律，即 Hub 特征、Linkage/Sibling Locality 特征、站点主题特征、Tunnel 特征。这些特征的挖掘和利用对主题信息搜索过程中的链接预测和页面过滤方面具有重要作用，从而将搜索集中于更有价值的区域。同时，通过集成网络社区资源形成索引，从而对所有社区资源进行搜索成为可能。有研究提出了一个基于博客的爬取、过滤、索引和搜索方法，同时在分析 Web 社区搜索资源分散特点的基础上，运用向量空间模型和相关性排序等技术设计了 Web 社区搜索引擎的体系结构，实现了一个 Web 社区搜索引擎系统。

互联网数据挖掘与搜索技术向跨媒体方向发展是必然趋势，且已经成为国内外的研究热点。SNS 网络社区中不同媒体形式的数据对象之间存在广泛关联，例如，网页中实体与实体的关联，标签与网页的关联，用户对网页、多媒体对象的评论推荐等。多媒体数据对象之间的联系为对象的分类聚类提供了独特的信息来源，这种方法被证明是行之有效的。对于关联数据对象之间关系数据稀疏的问题，可以通过潜在语义分析法在低维语义空间中表示数据对象来解决。这种降低维度的技术通过许多应用拓展了它的表现。由于潜在语义分析法仅考虑了两种数据对象之间并发的关联关系，所以其实际应用受到了很大限制。

基于 Agent 的分布式网络计算模式也为复杂的跨媒体网络信息搜索注入了新的活力。将 Agent 系统在解决分布式计算问题时所具有的智能性、异步性、移动性等特点结合到 Web 搜索中具有诸多优点。Agent 技术与 Web 数据库相结合，能够更加准确地描述用户需求，从而进行有效的 Web 挖掘。移动 Agent 计算模式在 Web 数据库、Web 挖掘、信息搜索等分布式应用系统中都得到了很好的应用。

互联网搜索技术从依赖关键词匹配的方式向基于深度语义分析和挖掘的精准搜索方向发展是必然趋势，且已经成为国内外的研究热点。一个有效的面向领域搜索方法必须要考虑以下问题：如何设计计算高效的分布式多模态数据挖掘和搜索算法，并且能够适应 Web 3.0 时代复杂的网络和终端环境，以及网络社区的动态变化；如何从跨媒体的信息中获得语义信息；如何基于跨媒体的特征信息对网络社区内具有关联性的数据进行组织和搜索。根据多 Agent 系统在分布式计算领域的研究成果可以发现，将 Agent 应用于网络社区内跨媒体数据的搜索与挖掘是可行的，但如何解决上述关键问题，目前的研究尚属空白。

SNS 网站的潜在商业力量就是其中积累的海量数据。传统搜索引擎通过挖掘网民的搜索行为来提供很多有商业价值的调查研究报告，而 SNS 网站则包含大量的用户数据以及比搜索引擎更为丰富的互动数据，具有巨大的商业价值。Facebook 计划利用其庞大无比的用户数据，创建全球最大的市场研究数据库，提供可实时反馈的市场调查服务。而要理解如此海量和复杂的数据并从中得到有用的知识，必然要借助有效的 SNS 数据挖掘和主题分析工具。各种数据挖掘和主题分析技术得到了广泛应用，如基于网站用户信息和用户间关系进行聚类分析以及基于分类器的垃圾信息过滤，都用到了数据挖掘和主题分析技术。可以预计数据挖掘和主题分析技术的市场应用将会越来越多。

随着应用的深入，对主题分析技术也提出了新的要求。首先，SNS 网站间的共享与合作在不断加深。Facebook 目前正在开发的 FacebookConnect 功能，使用户能够在其他站点上看到好友的活动，也能够向好友广播他们在其他站点上的活动。用户在 Facebook 上的发言、在 Frienfeed 上的动态以及在 GoogleReader 上的分享，都可以通过 Twitter 推送给该用户的朋友。同样用户也可以通过其 Facebook 和 Twitter，接收来自各个 SNS 好友的动态、分享、新闻和图片。MySpace、雅虎和谷歌近期也都公布了相似的计划。在这种趋势下，个人数据规模将急剧扩大、数据类型也更加复杂和多样化、数据也呈现出分布式的特点，因此需要开发能够适应海量、分布式、跨媒体的主题分析工具去理解这些数据的内容和主题。

其次，随着无线互联网应用的日渐广泛与深入，用户在手机等移动终端上的 SNS 应用需求越来越大。调查数据显示，有 40.9%的 SNS 用户期望可以在手机上使用 SNS 服务。目前我国手机上网用户规模已达 1.8 亿，SNS 手机应用有着良好的发展前景。适应无线网络应用的 Web 3.0 将实现不同终端的兼容，从 PC 到手机、PDA、机顶盒等移动终端上的应用指日可待。Web 3.0 模式下的 SNS 应用将彻底改变互联网的信息传播方式和商业应用模式。在这种趋势下，新的商业模式将越来越多地依赖于对用户数据的内容和主题的分析，实现对市场的准确预测以及对用户群的细分和精确定位。基于 SNS 模型的主题分析技术的市场应用十分