硕 士 学 位 论 文

基于多元表征的军事信息可信度研究

Research of Reliability on Multi-Representations Military Information

作 者 姓 名：张天宇

学科、 专业：计算机应用技术

学　　号：20809362

指 导 教 师：林鸿飞教授

完 成 日 期：2011. 05

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：_____

作 者 签 名 ：_____ 日期：_____年____月____日

摘　要

目前，互联网技术日新月异，Web2.0 的出现带给互联网一个全新的模式。在这个开放性、互动性、蒙面性和全民性的“可写可读互联网” 平台，Blog、BBS、TAG、SNS、RSS、WIKI 等社会化软件得以广泛应用。每个用户都拥有自己的博客、自己维护的维客、社会化书签或者播客，用户通过 RSS、标签或者邮件、即时通信工具（QQ、MSN）等方式连接到一起，社会化网络基本形成，鲜活的互联网呈现出信息来源多样化、内容庞杂化、更新快速化等突出特点。在集体智慧得以最大化体现的前提下，互联网给人们带来诸多便利和快乐，但同时也给人们带来了信息筛选、过滤和鉴别的难题。

孙子曰：兵者，国之大事，死生之地，存亡之道，不可不察也。而今，互联网军事信息情报也被列入“兵者”的行列，各国的军事情报研究部门开始日益重视针对互联网的情报搜集和网络监管，并且初见成效。但面对浩如烟海的军事信息，如何进行鉴别信息可信度的问题同样深深地困扰着情报工作者，大家使用信息聚合平台成功地解决了信息获取的难题。然而，开放的检索平台只能提供依据相关性和时效性得到的排序结果，信息的可信度却无法保证，一筹莫展的情报工作者不得不花费大量的精力去进行人工的鉴别。因此，如何从互联网及时、准确地获取可信度较高的军事信息成为军事研究领域所关注重点之一。

本论文针对互联网军事信息可信度特点，提出了多元表征军事信息可信度的方法。首先，在解析网页结构的基础上，采用定制规则的方法编写爬虫从互联网获取军事相关信息，并进行结构化存储；而后，对获取的信息使用 LDA 模型进行主题聚类；分别对信息可信度在信息源、信息相关评论和信息内容三方面的表征进行计算，依此来评价信息的可信度。对于信息源的可信度计算充分考虑了其三个主要构成部分：网站、页面和发布者，借鉴了 HITS 和链接可达性等思想分析源的可信度，并建立向量空间模型对结果进行融合，得到源表征的信息可信度计算结果；基于评论的可信度计算则依据情感词典分析了显性评论对信息可信度的表征，使用改进的概率主题模型得到了与信息相关的隐性评论，而后其对表征倾向性分析，综合得到评论对信息可信度的表征；基于内容的可信度计算是依据军事特征词典和网络流行语词典，计算内容对信息的可信度表征；最后，引入排序学习思想将独立表征的实验结果看作特征进行结果融合，对信息可信度的排序结果进行优化。实验证明，本论文提出多元表征信息可信度的方法有益于军事信息可信度的甄别，能很好地满足军事研究人员的需求，且多元表征融合比独立表征方法效果更好。

关键词：军事信息；可信度；排序学习；LDA 模型

Research of Reliability on Multi-Representations Military Information

Abstract

With the rapid growth of the internet, Web2.0 technology brought a brand new model for the internet. There were amount of social software applications on the internet which was open and interactive such as Blogs, BBS, TAG, SNS, RSS, WIKI and so on. Each Internet-user had their own blogs, WIKI, social tag and Podcast to contact with each other by software. Thanks to these applications being widely used, the Internet became more and more diversiform and complicated. On one hand the Internet made us feel convenient and happy but on other hand it brought us lots of troubles such as the reliability of the information on the Internet.

Sun Tzu said: The art of war is of vital importance to the State. It is a matter of life and death, a road either to safety or to ruin. Hence it is a subject of inquiry which can on no account be neglected. The Internet military information was of the art of war which was paid more attention by each country's military researchers. How to discriminate the reliable information from the mass was a big question for those researchers. The opening of the Internet not only brought convenience but it also cost us plenty of time to get the reliable information. Thus, how to get reliable military information became the primary research in military.

This paper proposed a multi-representations method to acquire reliable military information by considering the characteristics of the internet military information.First, we structurally stored the military-related information crawled from the internet. And then we got the themes which was clustered by LDA model and then got reliability each document of the same theme considering the information source, information content and the reviews. Among of the three considerations, information source consisted of website, web page and information publisher. The reliability of information source was based on the HITS and link-reachability algorithm. The text and military features was applied on acquiring reliability of information content. We analyzed the sentiment of the reviewers and applied improved probability model to got the got the reliability of reviews. At last, we ranked them by their reliability in the same theme according to learning to rank. The experimental results showed that the proposed method did military information acquiring and discriminating good to satisfy the demand of military researchers. It also showed that multi-representations method performance better than single-representation method.

Key Words： Military Information; Reliability; Learning Rank; LDA model

目　　录

1 绪论

1.1 研究的背景

近年互联网技术迅猛发展，2003 年，伴随着 Web2.0 这一新的互联网模式的诞生，Blog、BBS、TAG、SNS、RSS、WIKI 等字样开始不断冲击人们的眼球。个性图片、电子杂志、自制音频、创意视频、WIKI 式写作、博客的自由书写以及微博的狂潮，成就着互联网络的灿烂。SNS 圈子、即时通讯、社区、论坛、博客等带动了信息传播，引导话题的方向，甚至导致了潜在的舆论冲击。门户网站似乎已经不再是信息主要承载形式，平等、公开和交互的个性化、社会性交流方式改变着人们的网络生活样式。人们在畅快淋漓地享受网络带来的便利与愉悦的同时，对于网络信息的获取和鉴别开始困扰人们。当然，很多搜索聚合工具已经可以方便人们从不同的来源处获取信息，例如，谷歌、百度和搜狗等搜索引擎或者中华英才、赶集网等信息平台。人们开始利用“百度知道”、“腾讯问问”等进行常识查询；利用“维基百科”、“谷歌学术”等进行学术研究；利用“中华英才”、“智联招聘”等进行求职雇工；利用“淘宝网”“赶集网”等进行商品交易。有这样一个例子：淘宝网对于中国的网民基本都不会是一个陌生的名词，淘宝上的实物销售额一年超过 1000 亿。在一份研究表明：中国有 30%的网民在淘宝购物，而他们中又有 40%左右的人每周花费在挑选商品的时间在 14 小时以上。一位男士就有过这样的淘宝经历：他想在淘宝买部手机，结果返回结果包括 29534515 件宝贝；他又输入“CDMA”，结果仍有 94573 条；他继续输入“三星”，这次只剩下 7705 件了。继续吧，加入外观特征“直板”，结果仍有 4277 件，没办法了用淘宝的销量做个排序吧。一条条看下来，有的价格太高；有的检索到的标价很低，却是二手的或者是要求 100 台以上的批发价；同样的标价有的包邮，有的不包邮价格又不一样了；商品价格都一样了；有的商家信誉不好，有的商品评价不好……，一上午过去了，也没有找到所需的商品。大概很多上过淘宝的人都会有类似的经历吧。可见，即使是像淘宝网这样已经对商品进行了分类，也对商家进行了信誉评价的情况下，对于信息的筛选和甄别仍然是一个难题。那么，对于“专业人士”呢？哈工大刘挺老师就曾经在博客中讲述过自己的一个经历：一个雨天刘老师想打电话叫肯德基送餐，于是他上网找到了一个肯德基的 800 电话拨过去，电话里说“此电话尚未开通”。又找到哈工大附近那家肯德基的电话，打通了，一位老奶奶和蔼的声音响起：“小伙子，你打错了”。无奈，他拨打了 114 号码百事通，服务员说没有登记。继续搜索得到一条网上的留言说“肯德基不送餐”，不可能啊，再看留言的时间是 2007 年。历尽千辛万苦，他终于找到了 4008-823-823 这个电话，顺利完成了订餐。通过这个订餐的例子我们更容易感受信息可信度的重要性。因此，如

何在准确的前提下从海量信息源中快速发现某个特定领域的可用信息，将其用规范的形式加以描述，用合理的手段加以存储使用或者展示已经成为一个我们必须关注的问题。信息可信度的研究成果将使三方受益：对于用户来说，可以更方便地判定信息的可信程度；对于搜索引擎来说，可以有意识地抓取可信的信息，并在搜索结果排序时把可信的链接排在前面；对于网站制作者而言，可以更好地理解用户或搜索引擎判定信息可信度的方式，从而使自己制作的网站可信度更高。

同样的问题不仅仅困扰了普通的网民，在互联网信息的关注人群中还有一类特殊的群体——各国的军事情报部门。伴随着互联网络的诞生，军事情报部门的工作方式就发生了改变。他们不再倚重间谍等单一的方式获取情报，而借助互联网来公开搜集获取情报，或者对网络进行监管。2005 年美国中央情报局成立了公开信息中心负责搜集全球各个网站、论坛里军事信息；同年，英国也通过英国广播公司监测处获取全球网络信息甄选后提供给军事情报部门使用；2006 年，中国也正式将“网络战”写入了军方的教科书。2006 年 9 月 11 日《圣何塞信使报》报道，美国及中国部分的邻国正在探索一条窃取中国军事情报的“新路”——借助 Google Earth 或者 MSN Virtual Earth 等互联网图片搜索引擎，捕捉他们感兴趣的中国军事设施卫星图片。虽然，经过证实一些所谓的“情报”只是捕风捉影，但是美国国防情报中心主任特雷莎-希根斯指出：“我不觉得他们这样做毫无意义，因为对情报掌握得越多，你对目标的了解也就越深。”2010 年 7 月 26 日，号称“无国籍的新闻机构”——“维基解密”网站公布了 2004 年 1 月至 2009 年 12 月 91731 份阿富汗战争中鲜为人知的美国五角大厦（国防部）档案和战地情报机密文件，顿时全球哗然，《卫报》称之为“情报历史最大的泄露”。由此可见，各国的军事情报部门从互联网获益的同时，信息的搜集与甄别的难题也摆在大家面前。试想，花费着大量时间与精力，得到却是错误的甚至是虚假的信息，进而导致决策或者是指挥上的失误，其危害必定是无法估量的。因此，信息可信度研究，特别是军事领域信息可信的研究必将为各国的军事情报部门所重视，这也成为了本文研究重要动力。

1.2 研究的现状

信息可信度研究是一项新兴的兼具挑战性的研究课题，至今专门从事这方面研究的学者和相应的学术论文并不多，针对军事信息领域的更是寥寥无几。信息可信度的研究大体可以分为两个方向：一是信誉度或声誉度的研究，这方面研究的对象主要集中于网站和作者等信息发布者，例如文章[1-6]，在文献[1,2]提出了一种基于网络的网页搜索引擎思想，通过发现垃圾链接来评估网站可信度的方法；在文献[3,4]中对网站的可信度问题进行了研究，文献[3]中主要介绍了几种定量计算网站可信度的方法，文献[4]则将情

感和可视化研究引入到了新闻网站的评估中；文献[5]则提出了通过协商网络评价电子商场中卖家可信度的方法；在文献[6]中构建了基于评论用户可信度来改进在线声誉系统的信任模型。二是关于文本（包括新闻、博文、帖子、文章和电子邮件等）、图像、音频和视频等实体的真实性和精确性研究，而文本领域的研究又可以分为以下几个方向：博客[7]、电子邮件[8]、网页[9]和维基百科[10-15]等。在文献[7]中介绍了通过计算与新闻语料内容的相似性和结构的差异性来评估博客可信度的方法；在文献[8]中提出了通过识别实体和相关对象来评估邮件可信度的方法；文献[13]借鉴了 PageRank 和 HITS 思想，提出了依赖作者和修改者声誉度来评估文章的质量，并通过对文章修改和发表情况来动态调整作者的声誉度的方法。以上方法为本论文使用多元表征可信度提供了很好的参考借鉴。需要特别提出的是日本在可信分析领域的研究工作，其国家支持资助了两个研究组织 NICT（National Institute of Information and Communications Technology）以及 MIC（Ministry of Internal Affairs and Communications），其中，NICT 组织的 Sadao Kurohashi Susumu Akamine 等人[9]在对信息文档风格和表面特征研究的基础上构建名为 WISDOM（Web Information Sensibly and Discreetly Ordered and Marshalled）系统（如图 1.1）。

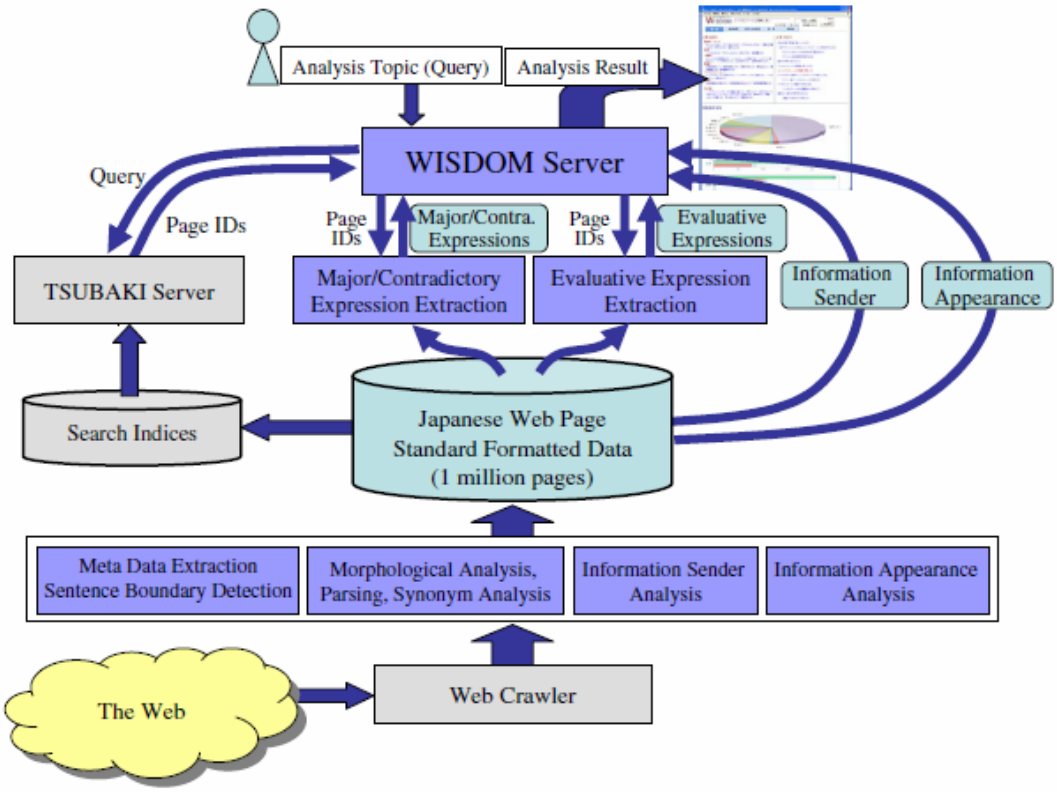

图 1.1　WISDOM 系统架构图

Fig.1. 1　The system architecture of WISDOM.

该系统通过评估信息评论的可信度和信息发送者可信度来评价网页内容可信度，具有较好效果。但该系统主要对信息的来源和评论等外部特征进行了评估，而忽略了对信息内容特征的可信度评估，而这一特征往往又是极为重要的。同时，该系统只给出了信息来源的具体情况分析和相关评论的统计，而没有对两者进行融合，自然也不能给出信息可信度的有序化结果；也有研究将信息的内容特征作为研究的主要对象，在文献[7]中就介绍了通过计算与新闻语料内容的相似性和结构的差异性来计算博客可信度的方法；该方法只能满足对某一类型信息的可信度评估，当信息来源和形式呈现出多样化的时候，则该方法不再适用，且可信度分析过程仅针对信息的某一方面特征，其结果的有效性不高。针对以上问题，本文提出了多元表征军事信息可信度的方法。

1.3 本文的工作

本文充分考虑目前信息可信度研究中的不足，立足于军事领域，以中文军事信息为研究对象，通过分析军事信息的网页结构、文档结构和表面特征等，提出了基于信息源、评论和内容的军事信息可信度多元表征方法，并在此基础上，将源、评论和内容的独立表征结果作为特征，采用排序学习模型进行了结果融合，给出了多元表征军事信息可信度结果。本文主要完成了三部分的工作：

一是建立了军事特征词典和网络流行语词典。根据研究需要，在对军事特征词语语法和构成等特点进行分析的基础上，基于词频统计与词法规则建立了军事特征词典，依靠网络资源搜集整理了网络流行语词典；

二是实验语料库的建立，即军事信息的爬取和处理。这部分主要是在对网页结构分析的基础上对军事信息的爬取。本文针对不同来源的信息，选取出 7 家军事网站、5 家军事论坛、3 家博客和新浪的 5 个军事研究员的个人博客作为爬取的对象，爬取了 2011 年 1 月 9 日至 11 日的军事文档共 4338 篇，对文档标题、来源、作者、是否原创、发表时间、相关 URL、评论信息和内容主体等内容及字体、字号及颜色等特征信息进行结构化存储。在此基础上对信息进行了预处理，引入军事特征词典和网络流行用语词典，利用分词工具完成信息（包括主题、内容和相关评论）切分等预处理工作，而后将信息划分为训练集和测试集两部分，引入 LDA 主题模型分别对其进行主题聚类；

三是信息的可信度表征，针对信息的源、评论和内容分别进行可信度表征，对于信息源的可信度表征充分考虑了其三个主要构成部分：网站、页面和发布者，借鉴了 HITS 思想和链接可达性等思想分析源的可信度，并建立向量空间模型对结果进行融合；基于评论的可信度表征对垃圾评论进行了过滤，而后依据规则对显性评论进行分析，最后，

使用改进的概率模型确定相关隐性评论，在分析汉民族语言特点的基础上使用情感本体库分析了隐性评论情感倾向性，将显性和隐性评论的分析结果融合作为基于评论的军事信息可信度表征结果。基于内容的可信度表征则引入军事特征词典和网络流行语词典进行特征计算来表征信息的可信度，而后简单地考虑了时效性等因素对结果进行调整；最后将独立表征结果作为特征，利用排序学习模型对独立表征结果进行融合，得到多元表征的军事信息可信度结果。

1.4 本文的结构

论文共由 5 章构成，较为详细地阐述了基于源的军事信息可信度表征、基于内容的军事信息可信度表征、基于评论的军事信息可信度表征；并介绍各方法实验设计和性能评估。具体章节安排如下：

第一章，绪论，综述了论文所研究内容的背景以及研究现状，介绍了本文研究的主要方法和论文的结构安排。

第二章，主要介绍信息可信度计算的相关技术以及本文所采用的评价方法。

第三章，详细介绍本研究所必需的军事特征词典的建立方法和实验语料爬取、结构化存储和进行预处理等过程。

第四章，详细介绍了多元表征军事信息可信度的方法，其中包括了基于源、评论和内容的军事信息可信度的具体算法实现以及实验设计与结果分析。

第五章，论文的总结，对本文研究的内容及主要工作进行了总结叙述，并就下一步的工作提出了几点思考。

2 相关知识和评价方法

2.1 信息可信度

可信度，蔡自兴，徐光佑等人在文献 [16]中较早从人工智能角度上给出了广义的定义：可信度（CF（H,E））是根据经验（E）对一个事物或者现象（H）为真的相信程度：

$$CF(H,E)=\begin{cases}\dfrac{p(H\mid E)-p(H)}{1-p(H)} & \text{当} p(H\mid E)\neq p(H)\\ 0 & \text{当} p(H\mid E)=p(H)\end{cases} \tag{2.1}$$

其中，p（H/E）表示在证据 E 出现的情况下现象 H 出现的概率，p（H）表示现象 H 出现的概率。信任通常表现出非对称性、可组合性和传递性等特征。具体说：非对称性表明甲信任乙，并不代表乙同样信任甲，或者说甲和乙相互信任的程度可能并不等量；可组合性是说如果甲、乙同时信任丙，则丙的可信任程度可以由甲、乙的信任度共同生成；传递性就是说甲如果信任乙，而乙又信任丙，则可以看甲同样信任丙，但这种信任的程度可能是不一致。当然，仅仅依据定义和特征分析来理解并实现信息的可信度的计算是十分困难的。

在文[17]中基于信任产生者的主观角度定义信息可信度为：“产生信任或信赖的程度”；而在文[18]中从研究对象的客观属性角度对信息可信度进行定义：信息的精确度或真实度。本论文认为将两者相结合更有利于对信息可信度的全面理解，同样也更有利于研究工作。在对信息可信度准确定义的基础上，可以将信息的可信度属性分为：诚信性（trustworthiness）、完整性（completeness）、专业性（expertise）、好意性（good-will）和时效性（timeliness）。分析信息可信度的定义和属性不难看出：信息可信度其实质是对信息质量的探究。在信息的达到量的积累后人们必定会要求质的飞跃，在寻求效率的过程中质量必将越来越为人们所重视。此时，人们将不再关注单纯的原始信息，而是需要那些可靠的有用的信息。那么，信息可信度的计算将会自然语言处理领域发挥至关重要的作用。

可信度的定义明确揭示了信息可信度就是对信息质量的研究。本论文通过分析发现信息不可信的产生是有规律可循的。一般来说造成不可信的因素主要有以下几点：

（1）录入错误：一方面是录入时的手误，例如：将 2010 年输入成 20010 年等；另一方面是输入法导致的错误，特别是中文录入表现明显，例如：在使用输入法联想功能时引起的多余文字录入形成语义改变，例如：录入“是”会联想出“不是”，如果不注

意就会把陈述句变成判断句。同时，不同输入法编码顺序不一致或者出现重码时也可能引起错误，例如：使用拼音输入法时将“阈值”输入为“域值”或“阀值”，使用五笔输入法时将“语义”输入成“主义”（编码同为 ygyq）；还可能因为信息传递过程中使用编码方式 （ASCII、GB2312、GBK、GB18030、Unicode、UCS 和 UTF 等）不一致等因素影响导致的乱码错误；如果使用扫描仪或者手写板时，还可能出现把字母“i”识别成字母“l”或者数字“1”等识别错误；

（2）作者不专业或者记忆有误：作者可能对某个专业领域缺少深入的研究，对于一些定义、术语出现了理解的偏差，也可能在一些重要的且复杂数据记忆上出现了错误，如顺序颠倒等问题，或者是在翻译的过程中出现了不准确的问题，例如：将“Maximum likelihood estimates”（最大似然估计或极大似然估计）译成“极度相似概率”。

（3）以讹传讹：主要是转帖或者分享了别人的不可信信息，这类情况也是极为常见的，在 Web2.0 时代转发与分享已经变得经常，而在这一活动仅是转发者一时兴起，在“转”与“分”的过程中缺乏必要鉴别，或者鉴别能力不足导致这个错误的传播；

（4）过时信息：可以理解为具有时效性的信息因为过时而变得不可用，例如：在战争结果时搜索到了战争初期发布的作战实力（参战 5300 人）和伤亡情况（有 72 人死亡），对于战争情况的统计已经没有意义。在日常生活中，如上文所提订餐电话的问题也是这样的。

（5）残缺信息：这是信息质量的一个重要评定标准，其缺失过程可能出现在信息发布或传递过程中。例如：对于一篇军事分析，前面多是一些事态或形势的说明，后面才是作者的分析和观点。如果缺少了前一部分，读者会觉得言之不实，缺乏可信性，如果缺少了后一部分，读者就会有言之无物的感觉，对于数据的堆砌丧失兴趣，只有完整的呈现才会易于且乐于被读者接受；

（6）恶意造假：这类情况在网络中屡见不鲜，其出现往往具有明显的主观性和强烈的目的性，是主观炮制错误信息试图从中获益的过程。有的只是单纯地为提高声誉引起关注，有的则是希望通过恶意欺骗的行为达到个人（组织、国家）某种目的。

对于以上信息不可信的成因，本文大致将其分为两类：一是客观成因（如 1-5 条），这些信息往往都不是主观行为导致的，对于这些信息可以通过定义规则、设计模型、比对专业网站、进行网络推理和建立常识库等方法进行评估；二是主观成因，这部分主要是指第 6 条原因所形成的不可信信息，信息的发布者通常是该领域内的专家，其会有意识且有能力掩盖自己的目的，信息可能会夹杂着大量真实信息或数据来蒙蔽读者或其它信息受众，试图让大家客观的认可信息的质量，这部分信息的甄别无疑具有较大的难度。如果是商业信息（如商品的买卖、二手交易等），可以通过对信息的发布平台和发布者

声誉等方法进行鉴别，也可以通过相关的评论或者是比对其它类似信息的差异程度来分析。但对于军事这个特定的领域，信息的提供者可能是某个国家或者组织时，就只能建立模型进行推理了，但推理的结果只能是一个参考的建议，更多时候需要人工借助更多的外部信息进行辨别。因此，本文的研究主要还仅仅只能针对第一类成因的不可信信息。

2.2 LDA 模型基本思想与应用

Latent Dirichlet Allocation（LDA）模型[19]是由 Blei 等人于 2003 提出的多层次主题产生式全概率图模型，其前提条件是“词袋（bag of word）”假设，即在忽略任何语法结构和词序关系的前提下将文档看作是独立词条的集合。LDA 模型的理论基础是基于极大似然估计（Expectation-Maximization （EM） Algorithm）、变分近似（Variational Inference）、吉布斯采样（Gibbs sampling）和贝叶斯网络（Bayesian Network）等经典算法思想，该模型被广泛应用于文本挖掘、信息检索等诸多文本研究的相关领域，经典应用是对文本数据的主题信息建模。

首先 LDA 模型是主题模型，所谓主题模型是指在计算机无法真正理解自然语言的情况下，提取出可以被人理解的，相对稳定的隐含语义结构，为大规模数据集中的文档寻找一个相对短的描述。一般包含了词、主题和文档三层结构，呈现出两种分布：一是主题-词的分布；二是文档-主题的分布。具体说就是：对于一个给定的文档集合，LDA 将每个文档表示为多个主题的集合，而每个主题符合一个多项式的分布，从而建立词之间的关联。换句话说，在 LDA 模型中全部主题被所有文档所共享；每个文档按各自特定的主题比例被描述。

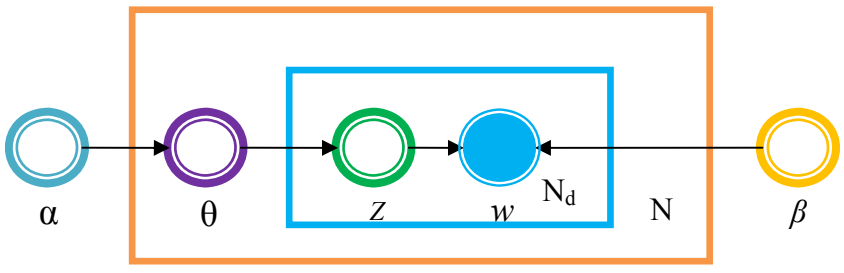

图 2.1 LDA 的图模型表示形式

Fig.2.1 Graphical model representation of LDA

LDA 模型在文本建模过程中还是生成模型，所谓生成模型是指可以随机生成可观测的数据，LDA 可以随机生成一篇由 N 个主题组成的文章。几乎所有讨论 LDA 的文章都将无一例外地引用了图 2.1，因为其是最形象直观 LDA 的模型样式。

其中，蓝色实心部分表示已知的可观察到的内容、数据或者值，而空心部分则是需要我们求解的隐含的或者潜在的变量，橙色大矩形表示反复抽取的主题分布，蓝色小矩形表示通过反复抽样产生文档的词。文档层的参数（α,β）共同确定 LDA 模型，α 反映了文档集中隐含主题间的相对强弱，β 代表了所有隐含主题自身的概率分布。θ 代表文档中各隐含主题的比重，z 表示文档分配在每个词上的隐含主题比重，w 是文档的词向量表式。N 为文档集中文档个数， N_d 表示该文档的词总数。其概率模型可以表示成：

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \qquad (2.2)$$

其中，p（$\theta|\alpha$）在 α 条件下产生 θ 的概率，p（$z_n|\theta$） 在 θ 条件下 z_n 产生的概率，p（$w_n|z_n$, β）是在（z_n, β）条件下产生 w_n 的概率。其生成过程形式如图 2.2：

1. 选择文档长度N，使N~Poisson(ξ)分布；
2. 选择主题发生的概率θ，使θ~Dirichlet(α)分布；
3. 对于由N个单词构成的文档d中每一个词：
   a) 选择主题z_n，z_n服从Multinomial(θ)多项分布。z_n代表当前选择的主题；
   b) 根据z_n条件下的多项分布p($w_n \mid z_n; \beta$)，选择w_n。

图 2.2　LDA 模型概率生成过程
Fig.2.2　Probability generating process of LDA

经历了从 LSA 到 PLSA 再到 LDA 的过程，文本建模思想得以逐步的完善，甚至于完美。LSA 通过向量空间的降维实现了潜在语义空间的挖掘，与 LSA 相较 LDA 中的实现方法则是通过文本数据对主题空间映射获得文本间的关系，具有高效的概率推理算法。换句话说，LDA 模型采用少量训练样本通过无监督方法即可实现对模型的训练，大规模的文本样本有助于优化模型模数，运算的时间复杂度远远小 LSA 模型，因此更适合处理大规模文本语料；与 PLSA 相比，LDA 内存结构更加清晰，其在文档-主题层即引入了狄利克雷分布，使得模型参数的数量保持在一个静态相对稳定的状态下，不会像 PLSA 一样随着语料库的扩大而增多。

在 LDA 的建模过程中，最关键是对参数的估计。其不再使用与训练数据直接联系的个体参数集合，而是将主题混合权重当作 k 维参数的潜在随机变量，同时采用了拉普拉斯近似、变分近似、传统马尔科夫链蒙地卡罗算法和期望扩散等方法获取待估参数值。本文利用比较常用的 EM 算法[20]进行参数估计，它主要分为两步：E（Expectation）_Step