专题报告

智慧城市背景下公众反馈大数据分析：概率主题建模的视角*

马宝君① 张楠②** 孙涛②
①北京邮电大学经济管理学院 北京 100876
②清华大学公共管理学院 北京 100084

摘 要：以智慧城市背景下公众意见反馈内容分析为目标，基于概率主题建模的潜在狄利克雷分配模型，提出了从大规模公众反馈信息文本中提取政府或政策制定者可能关注的潜在主题及讨论热度时序分析的初步方法框架；对某城市公众反馈网络平台2006－2013年间产生的实际数据进行了案例分析，展示和验证了分析结果及方法框架的有效性。

关键词：电子政务；公众反馈；大数据；智慧城市；概率主题建模

一、引言

随着网络技术的迅猛发展，公众通过各种网络平台参与公共事务讨论的渠道日益丰富[1-2]。与之相对应，政府部门也希望通过了解公众对于政策、行业的关注点、意见建议和反馈等信息，发现问题并解决问题。在Web2.0时代公众广泛参与的理念下，对上述信息进行有效分析并实现决策支撑是实现智慧城市管理的一个重要途径。

目前，由于网络信息量规模巨大，政府或政策制定者无法在短时间内阅读和了解公众通过各种渠道反馈的详细信息；同时，由于公众反映的城市建设过程中的问题角度多种多样，且随着时间的变化，问题也会随之演变，所体现出的公众关注及社会存在的问题也有所不同。因此，随着信息技术的发展及智慧城市建设的需要，帮助政府或政策制定者简单、方便、快捷、全面、有效地了解公众意见反馈的信息就显得格外重要[3]。

本文以智慧城市管理为宏观背景，以实现公众意见反馈内容有效分析为目标，基于概率主题建模的潜在狄利克雷分配模型，提出了从大规模公众反馈信息文本中提取政府或政策制定者可能关注的潜在主题信息及讨论热度时序分析的方法框架，并对某城市公众反馈网络平台2006－2013年产生的实际数据进行分析，展示了初步分析结果及方法框架的有效性。

本文的后续内容安排如下：第二节将介绍相关研究工作；第三节提出本文的研究方法框架，并详细介绍其中的潜在狄利克雷分配模型及方法框架的具体处理过程；第四节以某城市公众反馈网络平台上的公众真实反馈内容为案例进行分析；最后在第五节给出文章总结及未来的进一步研究方向。

二、相关工作

随着信息技术的发展，相关信息技术也逐步应用到公共管理与公共政策领域，用以在大规模数据背景下帮助更好地理解复杂的政策和管理问题。与之对应，

*基金项目：国家自然科学基金项目(71372044，71102010，71231002)、北京市哲学社会科学规划项目（12CSC014）以及清华大学文化传承创新项目（2012WHQN015）。
**通讯作者 收稿日期：2013-07-30 修回日期：2013-11-10

新近也产生了一门新的研究主题，被称为“政策信息学”[4]。

本研究所要使用的概率主题建模方法属于主题建模方法的一种，也是可以应用于政策信息学的重要基础方法。主题建模方法的目标是从大量的文本信息中发现并提取主题信息，最开始被提出和应用于信息检索领域[5]。

首先，通过间接对文档中包含的主题信息的处理，TF-IDF模式[6]及向量空间模型[7]提供了一种粗略的描述和建模文档内容和主题相似度的解决方案。这种模型和方法的缺点在于无法区分字面意思不相同而主题意义相同或相似的词语或文档。为了解决这一问题，学者们引入了奇异值分解（singular value decomposition）并提出了潜在语义分析方法（latent semantic analysis，LSA），将高维的文档词语空间转化映射成为低维的主题向量空间[8]。

接着，概率潜在语义分析方法（pLSA）被提出用以在概率框架下生成信息主题集合[9]。与LSA方法中主题表现为启发式的物理距离不同，pLSA中的主题表现为概率。由于不需要进行复杂度较高的奇异值分解，pLSA方法被成功应用于许多信息检索领域的大规模数据集的测试中。

潜在狄利克雷分配（latent dirichlet allocation，LDA）是一种有效的主题建模工具，被用来从文本信息中发现和提取表示文档的低维主题集合[10]。LDA模型使用贝叶斯概率框架进一步改进了pLSA模型和方法。在以上所有这些主题模型中，LDA模型的建模效果表现最好[10]，这也是本文基于LDA展开工作的原因。

三、研究框架

为了解决智慧城市背景下公众意见反馈内容的大数据分析问题，本研究应用了一种名为潜在狄利克雷分配的概率主题建模策略，设计了智慧城市背景下公众意见反馈分析的方法框架。在该框架中，LDA模型被用来从大量的公众生成的反馈文档信息中发现和标注相应的主题信息，以及发现主题间如何联系、主题如何随时间变化等问题。基于此，我们可以为政府或政策制定者提供他们感兴趣的经过滤的准确信息内容。

（一）潜在狄利克雷分配模型(LDA model)

潜在语义模型是一类从文本文档中提取潜在语义信息的有效方法[5, 8-10]，在这类方法中，潜在狄利克雷分配模型的表现最有效[10]。

LDA模型背后的直观假设很简单，即为每一个文档都显示了多个主题信息[5]。该模型是一种无监督的生成统计模型，目标是通过提出一种文档中词语生成的随机过程，找到文档的主题信息。LDA模型中假设所有的主题是在任何文档生成之前就已经确定。给定语料库中的任何文档，它的生成过程包含两个阶段。首先，随机选择符合狄利克雷随机分布的一个主题分布向量，用以确定该文档中的哪些主题最可能出现。然后，对于文档中出现的每一个词语，随机选择主题分布向量中的一个单独主题。为了真正生成该词语，我们随之使用选定主题下的条件概率。可以看到，文档中的每一个词语都是基于不同的随机选择的主题产生的。

总的来说，LDA模型不需要对文档做任何先验的标注或标签，潜在的主题信息是从对原始文档集合的分析中得到的。LDA模型可以帮助我们组织、分析和汇总大规模的文本信息内容，相比较使用人工标注完成以上工作是完全不可行的。

（二）研究过程

采用LDA模型分析公众意见反馈内容的具体过程描述参见图1。

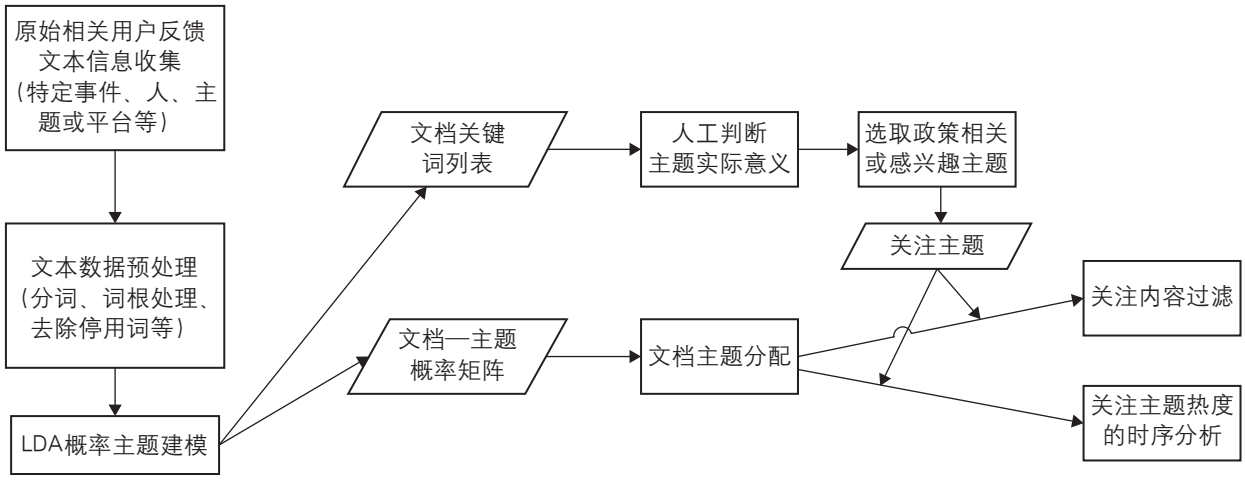

图1 采用LDA模型分析公众意见反馈内容的具体流程

首先，我们根据特定的事件、人、主题或具体平台收集原始相关文本数据信息。在此基础上，为了更深入细致地分析文本内容，需要对原始文本内容进行必要的预处理操作，例如分词[11]、词根处理[12]、去除停用词[13]等。完成数据预处理后，每一个文档都被表示为若干词语组成的向量。

其次，我们利用LDA进行概率主题建模。LDA模型可以对所有的文本向量进行训练和推理，然后提取出蕴含在这些文本数据中的潜在主题信息[5]。通过概率主题建模，我们可以得到以下两组有用的结果：

主题关键词列表——每个潜在主题下最相关的一些词语；

文档—主题概率矩阵——即每一行对应一个文档，每一列对应一个潜在主题，矩阵中的数值表示对应文档属于对应主题的概率值。

再次，我们使用得到的主题代表关键词列表，可以人工直观地判断相应主题的实际含义。相比较人工阅读全部的文档信息，由于一般的潜在主题数量不大（往往小于1000），人工阅读主题关键词一方面效率较高，另一方面也非常直观，效果也很好。通过判断潜在主题的实际含义，我们可以直接选取政府或者政策制定者关注的主题，也可以将有实际物理意义的主题关键词列表呈献给政府或政策制定者，让他们挑选感兴趣或关注的主题。

在此基础上，利用得到的文档—主题概率矩阵，我们可以按照最大概率原则，将每一个文档分配给隶属概率最大的一个或几个主题，表示该文档最有可能涉及到某个或某几个主题。如果需要在大规模的文本信息中筛选出和政府或政策制定者关注的主题最相关的文档集合，可以使用该阶段的方法将分配给关注主题的文档全部提取出来，从而完成关注内容的过滤。

最后，对于政府或政策制定者关注的某个或某些主题，可以采用累计所有文档在这些主题上的（最大）隶属概率来反映公众对于这些主题的讨论和关注热度；同时，在不同的时间段分别计算累计概率，可以实现关注主题热度的时序分析。

四、案例分析

（一）案例背景

本研究所选取的公众反馈网络平台是某城市纠正行

业不正之风办公室与市经济信息委员会共同建设的多渠道公众反馈数据平台，于2005年在政务门户网站上开通运行，主要通过电子政务手段促进公众对于智慧城市建设的参与，也为民主政治建设提供了典型实践。该平台试图发挥网络高效快捷、生动直观、互动性强等特点，24小时接收群众对城市政府部门、公共服务行业政风行风建设的咨询、意见建议和投诉举报，及时解决损害群众利益的不正之风问题。

平台开办八年来，累计受理网络信件24万余封，为群众解决实际问题16万余件，得到了社会各界的广泛参与和关注。目前，平台每天能够收到的信件都在百封以上，这些信件涉及到老百姓生活的方方面面。而对于信件中涉及的问题属于哪一种主题或问题分类，以及需要哪些部门来协调完成问题的解决，都需要相应的专家或顾问进行人工识别和判断。对于已有的公众反馈信件内容的主题、时序变化、关注热度的深入分析工作，目前由于数据量较大还没有深入开展。

（二）数据

本研究的案例数据来自某城市网络公众反馈平台系统。该系统上的注册用户可以留言发布咨询、建议或者投诉信息。研究样本涵盖了2006年5月23日至2013年4月12日间2517天的全部24万余封网络信件。本研究的分析重点是信件内容和信件标题，同时结合发信时间进行时序分析。

在获取全部数据的基础上，首先需要对数据进行预处理。我们发现，有一些信件内容是完全或基本相同的，我们使用文本匹配和近似匹配的方法将重复内容的信件删除，只保留其中的一条。此外，信件中存在大量的测试信件，即为工作人员测试系统时留下的，我们也通过信件标题和信件内容中都包含“测试”或“test”关键词的策略进行过滤筛选。通过以上处理，我们保留了197751封信件用于实际分析。

接下来，我们将信件标题和信件内容整合在一起作为每一封信件的文本信息，对每一条文本信息进行分词、去除停用词，将其表示成为词语的列表或向量，为使用LDA模型进行概率主题建模做好数据准备。在此过程中，由于信件中的内容基本全部为中文，我们使用了一个开源的中文分词软件包——“庖丁解牛[1]”来对中文的信件文本内容进行分词处理；去除停用词操作采用了Apache Lucene的SmartChineseAnalyzer类[2]中使用的中文停用词列表。此外，对于中文文本的处理与分析，本研究使用Java API和Apache Lucene API编程实现。

（三）结果与讨论

对于197751封公众反馈信件的文本内容，我们使用LDA模型进行了概率主题建模，选取的潜在主题数目分别为50、100、150、200和250。通过人工阅读不同主题数目下的主题关键词列表，可以发现主题数目为200个时的主题关键词列表的语义信息表现最好。限于篇幅，这里只在表1中展示了其中三个潜在主题的长度为10的关键词列表。

表1 案例中的三个潜在主题的关键词列表

主题192（公交线路）	主题148（环境污染）	主题136（医疗报销）
公交 线路	污染 环保	报销 医疗
增加 开通	空气 垃圾场	生育 医保
调整 增设	排放 严重	药费 住院
方向 换乘	污水 焚烧	定点 负担
终点 沿线	臭味 刺鼻	比例 公费
……	……	……

1 庖丁解牛中文分词软件包：http://code.google.com/p/paoding/。

2 SmartChineseAnalyzer类的介绍参见：http://lucene.apache.org/core/old_versioned_docs/versions/3_5_0/api/contrib-smartcn/org/apache/lucene/analysis/cn/smart/SmartChineseAnalyzer.html。

在该案例中，为了发现公众集中反映的突出问题，我们直接利用第三节第二部分中介绍的方法，对200个潜在主题在所有的197751封公众反馈信件文本上的隶属概率进行累加，可以得到每一个潜在主题在所有这段时间公众的关注热度，这种热度不仅表现在主题对应的信件数量上，而且更加准确地反映在与主题相关程度的准确信息量上。换言之，采用文档—主题隶属概率累加，不是粗略地将主题相关的每一封信件同样对待，而是准确地区分与主题的具体相关程度，从而更为准确地评估公众关注主题的实际热度。

接着，我们按照累加概率递减的顺序将200个潜在主题进行排列，并将其中没有实际语义含义的主题去除，就得到了公众意见反馈的主题信息列表。对于有实际语义含义的潜在主题，可以根据主题关键词列表归纳概括出相应的主题描述。表2中列出了从所有信件中提取的前10个有具体语义含义的主题信息。从表2中可以看到，交通、户籍、道路施工、住房、环境卫生以及医疗费报销等问题在最近几年一直受到公众的持续关注。

此外，我们还可以对每一个主题的讨论热度进行时序跟踪分析，以观察主题受关注热度随着时间变化的情况。图2展示了随机选取的五个公众关注主题（即子女落户、公交车服务、道路施工维修、夜间噪音扰民以及环境卫生脏乱差）的讨论热度随时间（季度）变化的情况。

通过进一步分析和比照，可以发现一些有意思的现象和结果，通常与政府政策、实际情况相对应。例如，公众对于“环境卫生脏乱差”问题的反馈最近几年来总体呈下降趋势，也反映出政府在文明城市建设中发挥了积极有效的作用；同时，该问题也基本呈现出夏季反映较多，冬季反映较少的趋势，这也和实际情况与直观感觉是一致的。此外，公众对于子女落户主题涉及的相关问题的反馈，在2010年的4－9月达到一个高峰，这个结果可能与那段时间之前不久该城市公布新的户籍政策有直接的关系。

总体来看，通过分析可以发现，本研究提出的公众意见反馈分析方法可以帮助政府或政策制定者从大规模的公众反馈信息中发现和提取集中和突出的问题，以及分析公众关注问题热度随时间变化而发生波动的趋势和情况。

表2 从案例中提取出的前10个公众最关注的主题信息

主题ID	累计隶属概率	占百分比	主题描述
Topic 192	2889.91	1.462%	公交线路、车站建议
Topic 13	2449.84	1.239%	公交间隔、车次、人多、区间车等
Topic 80	2245.82	1.136%	户口、结婚、孩子落户、户籍
Topic 87	1982.46	1.003%	交通路口、红绿灯、天桥、人行道、车流控制等设置
Topic 128	1783.27	0.902%	公交车司机、售票员、排队等服务评价和投诉
Topic 168	1718.55	0.869%	政府、城市、政策、制度发展建议
Topic 199	1701.21	0.860%	道路拥堵、积水、维修、施工等
Topic 154	1560.43	0.789%	对投诉意见要求处理、调查、解决、答复、解释
Topic 29	1537.14	0.777%	地铁线路、站点的建议与评价
Topic 134	1508.06	0.763%	城市社区、配套设施、轨道等建设规划和建议

id	季度
1	2006.5-9
2	2006.10-12
3	2007.1-3
4	2007.4-6
5	2007.7-9
6	2007.10-12
7	2008.1-3
8	2008.4-6
9	2008.7-9
10	2008.10-12
11	2009.1-3
12	2009.4-6
13	2009.7-9
14	2009.10-12
15	2010.1-3
16	2010.4-6
17	2010.7-9
18	2010.10-12
19	2011.1-3
20	2011.4-6
21	2011.7-9
22	2011.10-12
23	2012.1-3
24	2012.4-6
25	2012.7-9
26	2012.10-12
27	2013.1-4

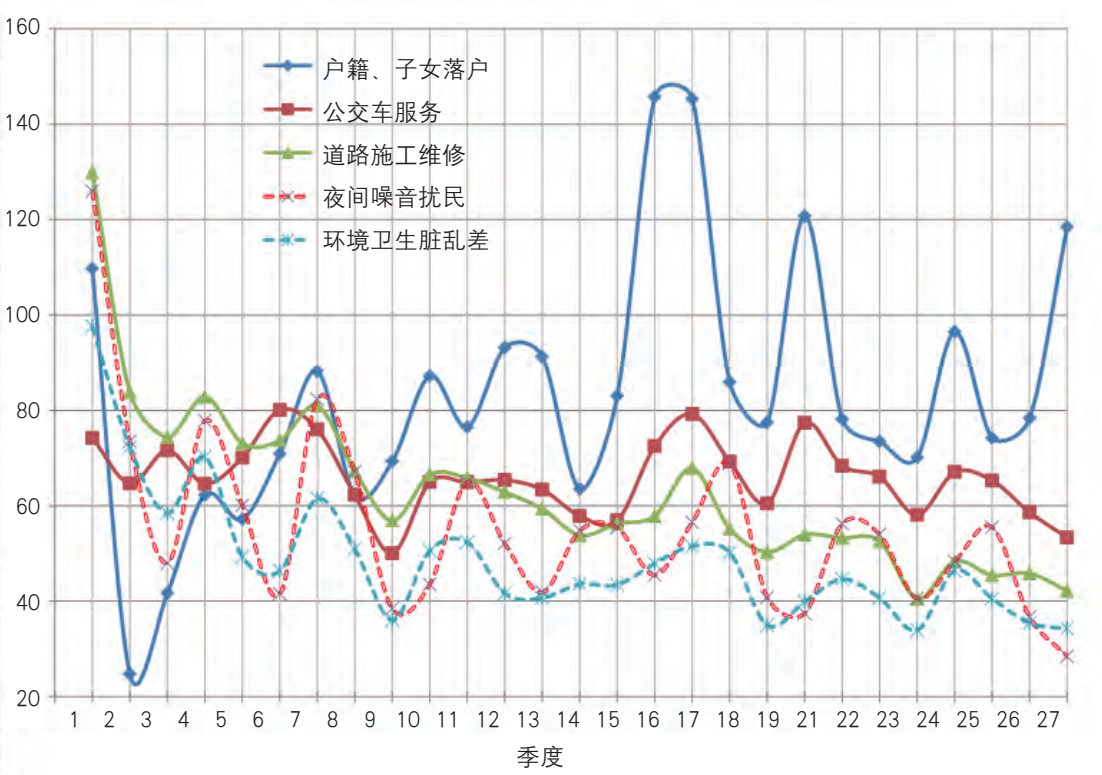

图2 五个公众关注主题的讨论热度随时间(季度)变化的情况

五、结语

本研究以智慧城市背景下公众意见反馈内容分析为目标，基于概率主题建模的潜在狄利克雷分配模型，提出了从大规模公众反馈信息文本中提取政府或政策制定者可能关注的潜在主题信息及讨论热度时序分析的方法框架，并通过对某城市网络公众反馈平台的实际数据进行案例分析，验证和展示了本研究所提出方法框架的有效性。

未来的研究可以在以下两个方面展开：一方面，公众反馈的意见态度和积极性对政府或政策制定者更为有用，因此后续工作可以结合文本情感分析的方法进行深入分析；另一方面，本研究涉及的平台也能够记录政府回复、解决公众提出问题的相关信息，后续研究可以利用这部分数据与公众反馈进行对比分析，以考察政府对于公众反馈的重视程度及应对措施的公众满意度等问题。

参考文献：

[1]黄军．网络行政论坛——政府形象塑造的新模式[J]．黔西南民族师范高等专科学校学报，2009(4)：12–15．

[2]徐双敏．公众参与政府绩效管理的现状与思考——以“民主评议政风行风工作”为例[J]．行政论坛，2009(5)：15–18．

[3]刘刚，詹建．公众反馈信息评价模型研究及实现[J]．软件，2012(7)：52–55．

[4]Johnston E，Kim Y．Introduction to the Special Issue on Policy Informatics[J]．The Innovation Journal：The Public Sector Innovation Journal，2011，16(1)：1–4．

[5]Blei D M．Probabilistic Topic Models[J]．Communications of the ACM，2012，55(4)：77–84．

[6]Jones K S．A Statistical Interpretation of Term

Specificity and its Application in Retrieval[J]. Journal of Documentation, 1972, 28(1): 11–21.

[7]Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing[J]. Communication of the ACM, 1975, 18(11): 613–620.

[8]Deerwester S, Dumais S T, Furnas G W, et al. Indexing by Latent Semantic Analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391–407.

[9]Hofmann T. Probabilistic Latent Semantic Analysis[C]// Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., 1999: 289–296.

[10]Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. The Journal of Machine Learning Research, 2003(3): 993–1022.

[11]Peña M, Bonatti L L, Nespor M, et al. Signal–Driven Computations in Speech Processing[J]. Science, 2002, 298(5593): 604–607.

[12]Hull D A. Stemming Algorithms: A Case Study for Detailed Evaluation[J]. Journal of the American Society for Information Science, 1996, 47(1): 70–84.

[13]Baeza–Yates R, Ribeiro–Neto B. Modern Information Retrieval[M]. New York: Addison–Wesley, 1999.

作者简介：

马宝君，男，博士，北京邮电大学经济管理学院助理教授，研究方向：网络搜索服务、电子商务与管理决策、数据挖掘与商务智能、政策信息学。

张楠，男，博士，清华大学公共管理学院助理教授，研究方向：电子政务与网络治理、智慧城市规划理论与方法、虚拟社会与网络文化。

EG资讯

中国软件评测中心发布2013年政府网站绩效评估结果

在2013年11月28日召开的“第十二届（2013）中国政府网站绩效评估结果发布会暨电子政务高峰论坛”上，中国软件评测中心张少彤副主任发布了最新的中国政府网站绩效评估结果，并指出了当前政府网站发展的几个主要特征。

张少彤副主任介绍说，2013年政府网站绩效评估更加关注网站的日常运维保障情况，加大了对公众关注度较高的重点服务的评估力度，并对利用新技术提升网站服务能力的情况进行了调查分析。

一、网站日常运维保障机制进一步完善

多数政府网站能够按照“国办函2011年40号”等文件的要求，定期开展自查自纠，及时发现并整改问题。部委、省、副省级、省会政府网站的首页链接全年可用性已经达到了99.1%，二级、三级页面链接的全年可用性分别达到了95.3%和81.2%，与2012年相比均有显著提升。各级政府网站加大了信息的发布力度，关闭了一批长期不更新的栏目。超过3个月不更新的栏目比例由2012年的48%下降至32%。

二、重点领域信息公开稳步推进

工业和信息化部、国土资源部、环境保护部、交通运输部、水利部、质检总局、上海、安徽、广东、福建、湖南、长沙、武汉、罗湖等政府网站，按照“国办函73号”文和“国办函100号”文的要求，建立了信息公开专题专栏，加强政策文件及解读、工作计划安排及进展等信息的及时全面公开。

三、服务丰富度持续提升，重点服务建设有待加强

多数政府网站能够按照党和国家的要求、社会公众的关注点建立服务专题，整合服务资源、丰富服务内容。同时，北京、湖南、湖北、佛山、长沙、柳州等部分网站围绕社会公众关注度高、办理量大的服务加强了重点服务建设，取得了较好的效果。

四、互动保障机制逐步完善，智能化交流平台建设开始起步

多数政府网站的互动交流水平持续提升，咨询答复时间进一步缩短，围绕社会热点的在线访谈和民意征集次数有所增加。部分网站的互动交流平台已经成为其履行职能的主要平台阵地，有效地支撑了业务工作的开展。此外，海关总署、国家林业局、济南、佛山、唐山、宿迁等部分政府网站基于历史资源建设互动知识库，开通了网上智能互动平台，提供更加及时、有效的交流，进一步提高了互动效率。

五、移动政务客户端建设任重道远

评估数据显示，我国移动政务终端的建设尚处于起步阶段，这主要表现在四个方面：一是移动政务客户端的建成开通率较低；二是安全隐患突出；三是服务内容单薄，更新维护有待增强；四是服务功能单一。