情报分析与研究

半监督的网络科技信息分类模型*

李传席　张智雄　刘建华　钱　力
(中国科学院文献情报中心　北京 100190)

摘要: 【目的】开放的网络科技信息网页内容之间区分度较小，传统基于规则和统计学习的方法无法满足网络科技信息网页分类的具体应用需求。【方法】通过深入分析网络科技信息主题网页的内容和结构，利用开放本体等资源实现领域特征的学习，构建半监督的网络科技信息分类模型。【结果】实验结果表明提出的方法在网络科技信息分类实验中的精度、召回率和 F1 值分别达到 0.9016、0.8756 和 0.8884，相比贝叶斯方法具有明显优势。【局限】该方法在应用到其他类别的网络科技信息分类时，仍然需要领域专家提供相关领域的核心种子特征。【结论】该方法可以满足网络科技信息深度加工的需求，实现有效的网络科技信息网页分类。

关键词: 网络科技信息　网络科技信息分类模型　开放资源
分类号: TP181　G356

1 引　言

互联网深入发展的同时也产生了大量的网络科技信息，采用高效的方法实现开放的网络科技信息的分类，有助于深度挖掘网络科技信息所蕴含的知识，满足科技信息加工处理的需求。网络科技信息资源监测系统[1]的目标是帮助战略情报研究人员全面及时地跟踪监测特定领域内一些重要科研机构发布的网络信息资源，网络科技信息资源监测可以洞察科技领域的发展态势，是文献情报研究的重要任务，而对网络科技信息的分类则是网络信息监测系统的重要环节之一。

为了深度挖掘网络科技信息资源中蕴含的知识，提升资源的利用价值，陈旭玲等[2]使用聚类学习算法分析科技文献中蕴含的创新研究方向，挖掘科技文献信息之间的关联。宋丹等[3]通过借鉴话题识别与跟踪方法中的技术，采用基于词包特征(BOW)的 K 近邻分类方法，利用文献之间的引用关系，实现科技主题的自动识别与跟踪。而刘勘等[4]则采用关键词技术聚类科技文献，利用特征词的词频等统计特征，构建文本向量，采用基于密度的方法实现科技文献的聚类，从目标科技文献中发现热点研究领域和研究方向，开展科技文献中知识挖掘的专题研究。贺亮等[5]利用科技文献元信息，采用 LDA 模型，通过自动分析科技文献的话题，实现相关科研领域的发展趋势和研究动态的挖掘。楚存坤等[6]采用模糊聚类技术，探索文献的自动分类技术。谢新洲等[7]讨论了网络信息资源的分类特征，并分析传统分类方法的适应性等问题。而刘建华等[8]则采用基于规则的方法对网络资源的标题进行分析识别，研究标题的来源和特征等方面在网络文本资源的标题快速识别分析上的效果。此外，网络科技信息的分析和处理研究中，王飞跃[9]阐述了面向大数据和开源信息环境下，科技信息向科技情报的实现过程中，也需要对科技信息进行深入的分析和处理，实现科技发展态势的预测和科技决策的制定与评估。刘云

收稿日期: 2014-05-20
收修改稿日期: 2014-06-24
*本文系中国科学院文献情报能力建设专项“网络科技信息自动监测系统二期建设”项目(项目编号:院1306)和国家“十二五”科技支撑计划课题“科技知识组织体系共享服务平台建设”(项目编号:2011BAH10B03)的研究成果之一。

等[10]则分析科技资源研究的相关手段和方法。

Qi 等[11]对网页信息的分类方法进行概述，并与基于文本的分类方法相比较,包括 KNN、Naive Bayes、SVM 等机器学习方法在网页分类上的效果，特征涉及词包特征 BOW、N-Gram 特征等在网页分类算法中的应用。Tsukada 等[12]给出 Naive Bayes、BayesNets、Trees、LinearSVM 等方法基于特定语料集训练后的分类效果，采用 BOW 和 N-Gram 作为特征的无监督学习方法却无法很好地实现网络科技信息的分类。与本文在实现科技网页分类时相似, Bartik[13]在网页分类时采用网页结构信息和网页内容相结合的方式。此外，Dumais 等[14]单独采用特征共现技术实现对网页文本的分类。

面对 Web 规模的网络科技信息，应采用有效的分类方法定位特定需求背景下的网页内容，加速信息获取的质量和效率，满足网络科技信息的深度分析和处理需求。但是由于这些网页的内容之间区分度较小，传统基于规则和统计学习的方法无法很好地对其进行分类。而与科技文献的分类方法相比，网络科技信息不具备结构化特性，分类方法不能得到很好的应用。本文通过深入分析网络科技信息网页的内容和结构，通过对 Web 资源进行了深度的剖析，利用开放本体等相关资源，构建一种半监督的网络科技信息分类方法，实现对网络科技信息的分类，并通过实验验证了本文方法在网络科技信息分类方面优于传统基于学习的分类方法。

2 半监督网络科技信息分类方法

图 1 是基于领域特征的半监督网络科技信息分类模型框架。该框架主要由网页数据块划分、数据预处理、特征选择与抽取、网页块隶属度计算、网页类别隶属度计算等部分构成。

2.1 领域特征的选择

根据特征重要性和表现形式的不同，将领域特征分成三类：核心特征词，由领域专家提供的领域词表组成，具有较强网页类别区分度；扩展特征词，包括采用特定方法对核心特征词进行扩展得到的特征词

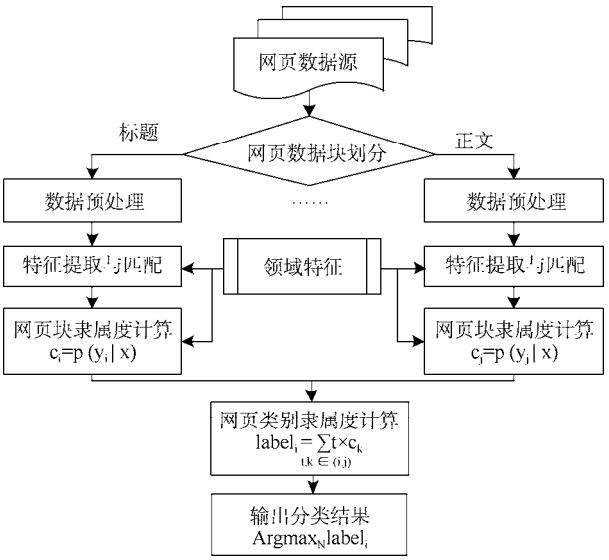

图 1 半监督网络科技信息分类模型框架

(见表 1)，对类别的标注起辅助作用，强度弱于核心特征；规则特征，即采用规则表达式形式呈现的领域特征。领域特征主要来源于以下三个方面:

(1) 核心特征词，由领域专家整理对类别具有高区分度的核心特征词。例如，对于研究报告，如果网页的标题中含有词汇特征“报告(Report)”，或者名词词组“某某报告(如 Technique Report)”，则此网页隶属于研究报告类别的强度很高，同时可以认为“报告(Report)”为核心特征词。

(2) 通过学习的方法结合不同的知识源，对领域核心词进行扩展学习，形成扩展特征词。考虑到领域专家整理的核心词虽然对于类别的区分度很高，但是存在一定的局限性，不能处理未登录词等特征，无法实现领域特征的全面广泛覆盖。因此，对核心特征从以下三个方面进行扩展:

①基于 WordNet[15]的同义词扩展。本文中采用 WordNet 对核心特征词进行语义扩展，获取同义词。例如，对于词汇 announce，可以扩展得到特征 announce denote, declare, annunciate, harbinger, foretell, herald。

②基于词典和叙词表的扩张。通过获取和分析词典与叙词表中的同义词，作为扩展领域特征，如在线词典①和在线叙词表②等。

③基于维基百科③的扩展。通过查询维基百科中含有的

①http://www.thefreedictionary.com.

②http://thesaurus.com.

③http://en.wikipedia.org/wiki/Policy.

与核心词相关的词条，对其进行分析并抽取相关特征，形成扩展领域特征。

(3) 对领域专家提供的领域科技文献语料进行整理、标注得到的结果，形成规则特征。

表 1　特征词统计

类别	核心词数	扩展词数	规则数
政策措施	226	454	15
机构调整	130	348	6
预算	227	431	14
重大成果	211	556	8
研究报告	162	501	11
统计评价	357	871	17
重大战略规划	257	491	3
重要战略声明	96	372	8
重要项目与计划	123	349	7

2.2　网页数据块划分与数据预处理

根据网页 HTML 结构信息，分析科技信息网页的结构和内容，获得网页中表达主要内容的网页数据块。根据实际需要选取网页中的标题块(Title)和正文块(Body)两部分作为有效信息内容块。

网页标题块和正文块的预处理步骤主要包括：去除网页块中的无关噪声信息，如 HTML 标签、样式、脚本等内容；对显示文本内容进行规范化处理，包括 Stem、Lemmatize 等。预处理步骤完成后，获得用于匹配领域特征的规范化网页文本块。

2.3　特征抽取与匹配

根据领域特征的内容，对数据预处理之后得到的规范化网页文本块进行分析与处理，抽取与匹配其中所包含的特征，形成网页内容的特征项表示，并提供给网页块隶属度计算组件。特征抽取与匹配的过程如图 2 所示，分别抽取网页块的原始表示特征(Raw)、Lemmatize 特征、词根特征(Stem)和规则特征(Regular Expression)等形式。

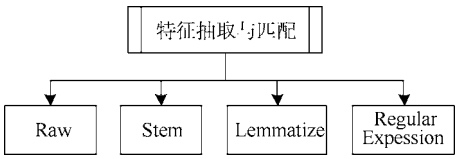

图 2　特征抽取与匹配过程

①信息监测服务平台 http://stm.las.ac.cn/STMonitor.

2.4　网页块隶属度计算与网页类别隶属度计算

根据网页的特征及领域特点，本文提出并构建如公式(1)、公式(2)和公式(3)所示的网页分类模型:

$$\mathrm{LabelN} = \mathrm{Arg\,max}_{N}\ \mathrm{label}_i, i \in 1,2,\cdots,c \tag{1}$$

$$\mathrm{label}_i = \sum_{t} t \times p(y_i \mid x) \tag{2}$$

$$p(y_i \mid x) = \frac{\sum_{i=1\cdots l}\ \sum_{j=1\cdots n} x_{ij} \times w_i}{\mathrm{Arg\,max}_{i=1,\cdots,c}\ p(y_i \mid x)} \tag{3}$$

LabelN 表示根据公式(2)和(3)计算得到网页的类别隶属值 Label_i 后，根据实际应用需求返回分值位于前 N 的主题网页类别。t 表示网页中不同网页块(网页标题、网页正文等)对于类别辨析力差异度，在此取 $t \in \{t_{title}, t_{body}\}$，$t_{title}, t_{body} \in (0,1]$, $t_{title} > t_{body}$，在应用中根据网页中不同块的重要性进行相应的扩展与调整。

$p(y_i|x)$ 表示网页 x 属于类别 y_i 的隶属值，x_{ij} 表示领域特征中属于类 i 的特征 j。w_i 表示第 i 类特征的权重值，用于度量特征的重要性，$i \in \{1,2,3\}$，$w_1, w_2, w_3 \in (0,1]$, $w_1 > w_2 > w_3$，w_1 为核心特征词的特征权重，w_2 为规则特征的权重，w_3 为扩展特征词的特征权重。在实验中，l 取值为 3，分别用于表示核心特征词、规则特征和扩展特征词。$\mathrm{Arg\,max}_{i=1,\cdots,c}\ p(y_i \mid x)$ 为归一化变量，用于将函数 $p(y_i|x)$ 的值转化为[0,1]之间的一个数值。

根据网络科技信息标注的需求，分类模型最终计算出网页属于每一类别的隶属值，形式为：$\{p(y_i) = [0,1]\}$，其中 i=1,…,9，y_1 表示政策措施；y_2 表示机构调整；y_3 表示预算；y_4 表示重大成果；y_5 表示研究报告；y_6 表示统计评价；y_7 表示重大战略规划；y_8 表示重要战略声明；y_9 表示重要项目与计划。

3　网络科技信息网页分类结果分析

选取网络科技信息监测服务平台①中的 900 条数据(5/01/2012-10/01/2012)，由相关领域专家进行标注，得到的结果如表 2 所示。

实验过程中，根据实际应用需求，抽取网页中的标题部分(Title)与正文部分(Body)，过滤去除网页中

表 2　评估样本数据量

类别	样本数
政策措施	100
机构调整	100
预算	100
重大成果	100
研究报告	100
统计评价	100
重大战略规划	100
重要战略声明	100
重要项目与计划	100
合计	900

与目标内容无关的信息。实验参数设置如下:

网页中不同网页块的参数重要性根据经验值进行设置。参数 t_{title}=1 表示网页中的标题对于信息的类别具有较高的辨析度，而 t_{body}=0.4 表示相对于标题内容而言，正文部分特征具有的辨析程度较小。而用于度量特征重要性的参数 w_1, w_2, w_3 分别设置为 1、0.8 和 0.6。

在实验结果的评估上，参数 N 取值为 2，表示同一个网页，可以同时属于一个或两个类别(主类别和候选类别)，对于属于两个类别的网页，由领域专家判断分类是否合理。算法的实验结果如表 3 所示:

表 3　基于领域特征的分类方法的结果

类别	Precision	Recall	F1 Score
政策措施	0.8598	0.9200	0.8889
机构调整	0.8065	0.7500	0.7772
预算	0.8713	0.8800	0.8756
重大成果	0.9326	0.8300	0.8783
研究报告	0.9468	0.8900	0.9175
统计评价	0.9592	0.9400	0.9495
重大战略规划	0.9192	0.9100	0.9146
重要战略声明	0.9474	0.9000	0.9231
重要项目与计划	0.8776	0.8600	0.8687
平均值	0.9016	0.8756	0.8884

为了更好地说明本文方法的效果，与贝叶斯方法(Naive Bayes)[16]进行了比较，训练二元分 S 类器对网页内容进行分类，训练过程采用留一交叉验证策略。特征包括 BOW、POS 和 Bigram，相应的领域特征也加入到贝叶斯方法的训练过程中，实验结果的评估采用三种度量指标，包括 Precision、Recall、F1 Score，如图 3 所示:

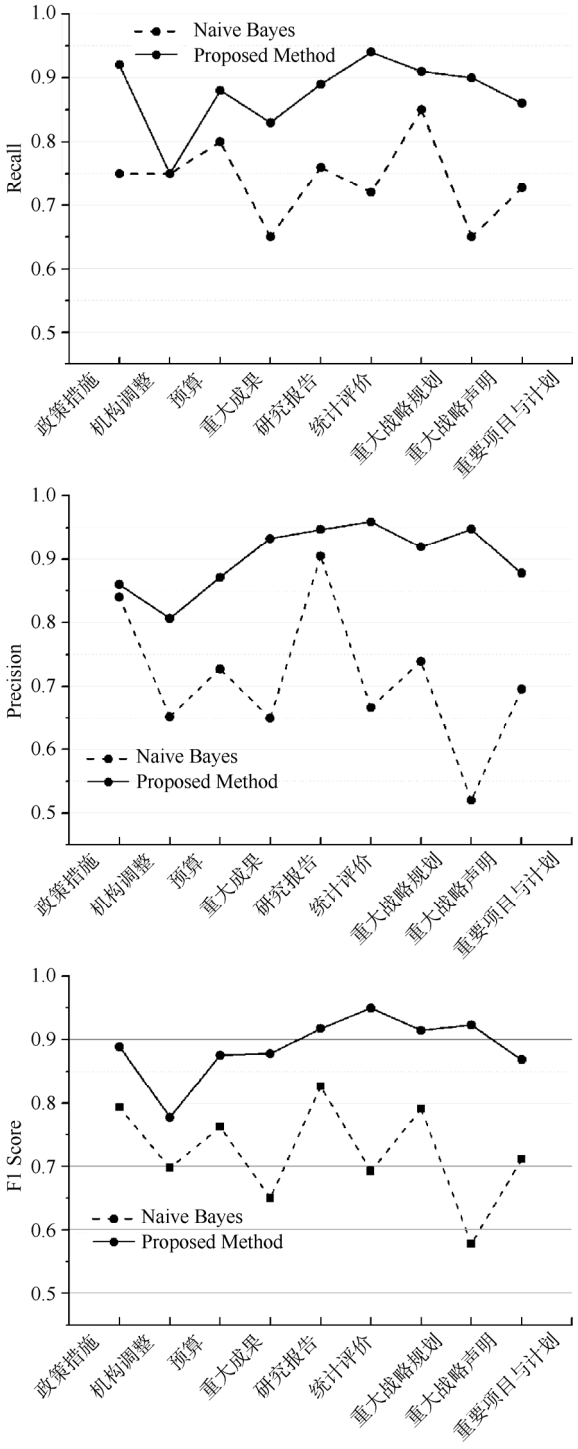

图 3　贝叶斯分类方法与本文方法的结果比较

从图 3 的结果可以看出，虽然在贝叶斯方法中加入了领域特征，但本文提出的分类模型在科技领域网页分类结果上明显优于采用贝叶斯方法得到的效果。

此外，经过领域专家整理出的背景知识在区分领域类别时具有更重要的作用，这也与 Rajaraman 等[17]的观点一致，即领域背景知识，例如关键词、短语以及规则等，对于区分领域内容具有重要的特征作用。通过检查误分类的网页发现，利用领域词典和叙词表对人工选取的领域特征进行扩展，可以弥补基于学习的方法在样本标注方面的缺点，能够发挥现有领域资源的作用。

另外，通过分析分类结果可以发现，一些启发式的专家经验规则对于提升分类效果起到积极的作用。与 Kan 等[18]中描述一致，URL 信息也只可以作为网页分类的特征使用，例如，URL 地址的目录信息，具有相同父目录的网页资源通常在类别上相同或相近：美国宇航局的网站(http://www.nasa.gov)，在 URL 目录 news/budget/下分布的信息多为预算或与其相关的信息。

4　总结与展望

网络科技信息资源监测是洞察科技领域的发展态势和文献情报研究的重要任务，有助于战略情报研究人员全面及时地跟踪监测特定领域内重要科研机构发布的网络信息资源，实现科技信息的深度分析，网络科技信息分类作为网络科技信息资源监测系统的组成部分，是实现网络科技信息加工分析的重要基础。由于科技信息网页的内容之间区分度较小，导致传统基于规则和统计学习的方法无法有效地对其分类。本文通过深入分析网络科技信息主题网页的内容和结构，结合领域专家知识，挖掘领域类别中含有的不同形式和类型的特征，利用已有的开放本体等相关领域主题资源，构建一种半监督的网络科技信息分类模型，实现对网络科技信息的分类标注。最后，通过与贝叶斯分类方法进行比较，验证本文方法在网络科技信息分类上的有效性。

参考文献：

[1] 张智雄，刘建华，邹益民，等．网络科技信息自动监测服务系统的建设[J]．科研信息化技术与应用，2013，4(2)：9-17．(Zhang Zhixiong, Liu Jianhua, Zou Yimin, et al. Implementation of Automatic Monitoring System for Science and Technology Information on the Web [J]. E-Science Technology & Application, 2013, 4(2): 9-17.)

[2] 陈旭玲，楼佩煌．改进层次聚类算法在文献分析中的应用[J]．数值计算与计算机应用，2009，30(4)：277-287．(Chen Xuling, Lou Peihuang. The Application of Improved Hierarchical Clustering Algorithm to Analyze Literature [J]. Journal on Numerical Methods and Computer Applications, 2009, 30(4): 277-287.)

[3] 宋丹，吴晨，薛德军，等．基于 KNN 的科技主题跟踪[C]．见：第五届全国信息检索学术会议论文集．2009．(Song Dan, Wu Chen, Xue Dejun, et al. Scientific Subject Tracking Based on KNN Algorithm [C]. In: Proceedings of the 5th China Conference on Information Retrieval. 2009.)

[4] 刘勘，周丽红，陈譞．基于关键词的科技文献聚类研究[J]．图书情报工作，2012，56(4)：6-11．(Liu Kan, Zhou Lihong, Chen Xuan. A New Clustering Algorithm for Scientific Literature Based on Keywords [J]. Library and Information Service, 2012, 56(4): 6-11.)

[5] 贺亮，李芳．基于话题模型的科技文献话题发现和趋势分析[J]．中文信息学报，2012，26(2)：109-115．(He Liang, Li Fang. Topic Discovery and Trend Analysis in Scientific Literature Based on Topic Model [J]. Journal of Chinese Information Processing, 2012, 26(2): 109-115.)

[6] 楚存坤，李韬．模糊聚类技术在文献自动分类系统中的应用[J]．现代情报，2009，29(9)：166-168，172．(Chu Cunkun, Li Tao. Application of Fuzzy Clustering Technology in Literature Automatic Classification System [J]. Journal of Modern Information, 2009, 29(9): 166-168, 172.)

[7] 谢新洲，金学慧，张婧，等．网络信息资源分类研究述评[J]．情报杂志，2012，31(2)：141-147．(Xie Xinzhou, Jin Xuehui, Zhang Jing, et al. Review of Network Information Resource Classification [J]. Journal of Intelligence, 2012, 31(2): 141-147.)

[8] 刘建华，张智雄，谢靖，等．基于规则的网络文本资源标题快速自动识别方法[J]．现代图书情报技术，2011(6)：27-31．(Liu Jianhua, Zhang Zhixiong, Xie Jing, et al. Automatic Identify Title of Web Text Resource Based on Rules [J]. New Technology of Library and Information Service, 2011(6): 27-31.)

[9] 王飞跃．知识产生方式和科技决策支撑的重大变革——面向大数据和开源信息的科技态势解析与决策服务[J]．中国科学院院刊，2012，27(5)：527-537．(Wang Feiyue. Decision Service and Academic Analytics for Development of S&T Based on Open Source Intelligence and Big Data [J]. Bulletin of the Chinese Academy of Sciences, 2012, 27(5): 527-537.)

[10] 刘云，王小黎，樊威．国际科技资源监测与服务体系构建

[J]. 科学学与科学技术管理, 2012, 33(8): 5-11. (Liu Yun, Wang Xiaoli, Fan Wei. Construction of the International S & T Resources Monitoring System [J]. Science of Science and Management of S. & T., 2012, 33(8): 5-11.)

[11] Qi X, Davison B D. Web Page Classification: Features and Algorithms [J]. ACM Computing Surveys, 2009, 41(2): Article No. 12.

[12] Tsukada M, Washio T, Motoda H. Automatic Web-Page Classification by Using Machine Learning Methods [C]. In: Proceedings of the 1st Asia-Pacific Conference on Web Intelligence: Research and Development. Springer, 2001, 2198: 303-313.

[13] Bartik V. Text-Based Web Page Classification with Use of Visual Information [C]. In: Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining. 2010: 416-420.

[14] Dumais S, Platt J, Heckerman D, et al. Inductive Learning Algorithms and Representations for Text Categorization [C]. In: Proceedings of the 17th International Conference on Information and Knowledge Management. ACM, 1998.

[15] Miller G A. WordNet: A Lexical Database for English [J]. Communications of the ACM, 1995, 38(11): 39-41.

[16] Hall M, Frank E, Holmes G, et al. The Weka Data Mining Software: An Update [J]. SIGKDD Explorations, 2009, 11(1): 10-18.

[17] Rajaraman A, Ullman J D. Mining of Massive Datasets [M]. Cambridge: Cambridge University Press, 2011.

[18] Kan M, Thi H O N. Fast Web Page Classification Using URL Features [C]. In: Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management. ACM, 2005: 325-326.

作者贡献声明：

李传席，张智雄：提出研究问题，设计研究框架；
李传席：研究方法的设计和实现，以及论文的撰写；
刘建华：提供部分实验数据和研究思路的讨论；
钱力：参与实验过程的设计与分析。

(通讯作者：李传席 E-mail: lichuanxi@mail.las.ac.cn)

A Semi-supervised Web Scientific and Technical Information Classification Model

Li Chuanxi Zhang Zhixiong Liu Jianhua Qian Li
(National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: [**Objective**] Considering the difference of open Web scientific and techical information is minor, general rule-based and statistical learning methods cannot classify the information effectively for the practical application demands. [**Methods**] By analyzing the content and structure of Web pages, and utilizing the open resources (such as domain Ontology and thesaurus etc.) to perform the self-learning of domain features, this paper proposes a semi-supervised classification model of scientific and technical information. [**Results**] The experiment results show that the proposed method achieves the precision of 0.9016, recall of 0.8756 and F1 score of 0.8884 respectively, which are superior to Naive Bayes classification. [**Limitations**] Applying the proposed method to new domain, the domain seed features need be supplied still. [**Conclusions**] The proposed method can classify the scientific and technical information effectively and satisfy the demand of the information deep analysis and process.

Keywords: Web scientific and technical information Scientific and technical information classification model Open resources