北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis

面向自动文摘的主题划分方法

童毅见　唐慧丰†

解放军外国语学院，洛阳 471003; † 通信作者, E-mail: tanghuifengom@126.com

摘要　对当前主题划分方法进行了分类，对主题划分算法 TextSegFault(TSF)作了相关改进，根据文本的类型，从 TSF 算法和改进的 TSF 算法中选择其一来进行主题划分，以适应自动文摘任务的需要。实验结果表明，引入本文的主题划分方法能有效地解决传统自动文摘方法造成的主题确实和主要主题冗余的现象，使文摘的结构平衡化。

关键词　主题划分；自动文摘; TSF 算法

中图分类号　TP391

Topic Partition for Automatic summarization

TONG Yijian, TANG Huifeng†

SPLA University of Foreign Language, Luoyang 471003; † Corresponding author, E-mail: tanghuifengom@126.com

Abstract　Current topic partition algorithm was summarized and classified. The authors improved one of the most effective topic partition algorithm—the TextSegFault (TSF) algorithm, and used TSF or its variant to partition topics based on the type of the text in order to meet the need of automatic summarization. Results show that the proposed method can help avoid the loss of minor topic or topic redundancy brought about by using traditional ways in automatic summarization, thus lead to the balanced structure of the summary.

Key words　topic partition; automatic summarization; TSF algorithm

1　主题与主题划分

1.1　相关概念

“主题”一词在很多话语分析的文献中均有出现，然而至今鲜有确切的关于主题的定义。Labadie 等[1]认为主题是会话或讨论的主要问题，而从语言学的角度来看，可以认为主题是正在讨论的命题。傅间莲等[2]认为所谓“主题”，是介于篇章与段落之间的一个语言单位，一个主题表达或阐述一个相对独立的意义或话题，从形式上由文章的若干个相邻自然段组成，各个主题相邻构成整个篇章。这个定义把主题看作是一个语言单位，并将其表达形式限定为连续段落集，有些欠缺考证。Brown 等[3]指出“主题”是用来描述一个话语片段所表达内容的一种直观方式。

主题划分，顾名思义，就是将一个含有多个主题的话语(在本文中以文本方式体现)切分成单个主题。Reynar[4]认为，作者在写作前，会在脑海中收集一些没有连接的主题，在写作过程中为了保证文本的流畅，会有意无意的设置一些主题边界。Brown 等也建议既然“主题”的定义比较困难，应该将精力集中在如何描述“主题切换(topic shift)”这个概念上来。主题划分的任务就是寻找这样的“主题切换”。实际上，主题和语言单位(如单词、句子、段落)之间并没有严格的对应关系，一个句子可能表达一个以上的主题，但是近似的，可以认为每个句子独立属于一个主题。

主题划分可以分为层次划分和线性划分，线性划分将文本切分成连续片段组成的线性序列，而层次划分则充分利用文本结构信息，如章节、子标题

收稿日期: 2012–06–06; 修回日期: 2012–08–30; 网络出版时间: 2012-10-26 17:49
网络出版地址: http://www.cnki.net/kcms/detail/11.2442.N.20121026.1749.014.html

等来进行主题划分[5]。尽管层次划分能取得很好的效果，但是大部分文章都不具有可利用的组织结构，因此这种方法的适用面较窄，本文主要讨论的是线性划分方法。从切分结果来看，线性切分还可以进一步分为连续切分和非连续切分，连续切分认为主题是连续的语言单位，而非连续切分则认为在一些写作风格比较灵活的文档中主题也可能是不连续的，例如第一段和第三段描述一个主题，而第二段描述另一个主题。

1.2 主题划分的意义

在文本，尤其是一些长文本中，其讨论的话题往往不止一个。最直观的一个例子是学术论文，通常一篇学术论文根据其章节划分，可以对应到不同的主题。Hearst[6]分析了一个包含 21 个自然段的学术文章《Stargazers》，并认为其可以分为 9 个子主题。

对自动文摘而言，如果采用传统的基于句子重要度从高到低抽取的方法，很容易造成对次要主题的遗漏或忽略，并且容易导致冗余现象的产生。自动文摘根据其目的可以分为面向主题、基于主题等，面向主题的自动文摘通常需要剔除次要主题，而基于主题的自动文摘则要求文摘系统对文本各主题及其联系有所把握，确保文摘能全面地反映文本内容，使摘要信息量最大化。无论是哪一种任务，主题划分都能起到重要的作用。此外，对于查询指导的个性化文摘，主题划分还能反映文本与查询之间的相关性，大大降低文摘中冗余句的数量。

除了自动摘要领域，文本切分在信息检索、文本理解、冗余消解、语言建模及文本导航等方面都有较大的作用。以信息检索领域为例，通常通过检索系统可以返回与查询相关的文本，但是该返回文本的多少内容呢？以视频检索为例，一个长达 4 小时的视频，其中与用户查询相关的只有 2 分钟，如果检索系统返回整个视频，那么用户将不得不看完许多与自己查询无关的片段才能定位到自己需要的信息。传统的信息检索方法，返回的是全文本或全视频，而通过主题划分之后，返回结果中可以只呈现和查询相关的主题片段，通过这种方式可以提高信息获取效率，对于长文本、音频、视频检索而言，这种作用尤为明显。

2 主题划分方法

关于主题划分方法，学者们也做过相关的分类。Labadie 等[1]将文本切分方法分为相似度切分法、作图切分法和词汇链切分法。Beeferman 等[7]将主题划分方法分为基于词汇连贯的方法、利用决策树融合多特征的方法及由 TDT 实验研究发展而来的方法等。

以上划分方法依据不同，略显混乱。本文拟将现有的主题划分方法划分为 3 类:

第一类是基于词汇连贯理论的方法。最初的主题划分研究多采用此法，也是目前主题划分领域研究最多，使用最广的方法。Halliday 等[8]指出词项的重复，尤其是带有内容的词项重复，导致文本连贯。该方法认为同一个主题内部，主题词会重复出现，并且与该主题词相关的内容词(除去功能词和少数频率极低的非功能词)也会大量出现。典型的算法包括 TextTiling[9]、C99[10]、dotplotting[11]等。

Hearst 等[9]提出的 TextTiling 主题划分方法被视作主题划分领域最经典的算法，很多关于主题划分的研究都是受到 TextTiling 算法的启示。该算法假设主题是由连续的句子集组成，首先在每两个句子之间设置位置点，并使用可滑动窗口，对每个位置点建立一个前驱块和一个后继块，块的大小与窗口大小一致。然后将句子用向量空间模型(VSM)表示，计算每个点前后块之间的相似度，利用作图法将结果表示出来，通过波峰和波谷来划分主题的边界。

基于词汇连贯理论的方法建立在词汇重复的基础上，具有一定的语种独立性，基于该方法的算法在主题划分上取得了很好的效果，但是该方法也存在一些明显的不足。第一，该方法极少使用句法信息，在自然语言中，一个名词是某个动词的主语与它是某个动词的宾语，意义完全不同。第二，文本风格影响算法效果，在法语中，一个词语重复出现是不雅观的，那么基于词频的方法此时效果不佳。关于这一点，有人试图通过引入同义词词典，如 WordNet 来解决，从一定程度上改善了切分结果。

第二类是融合特定语言现象和文本特征的方法。 Chafe[12]认为讲话者从一个焦点到另一个焦点(或从一个思想到另一个思想)就必然有一些点，在这个点上，时间、空间、人物配置和事件结构都发生了或多或少的变化，在某些点这些变化最大，呈现出来的就是明显的边界。这些变化很多时候就反应在特定的语言现象上。

Beeferman[7]利用了主题性特征和提示词特征，

根据标记的训练文本, 建立了一个指数模型来识别文本边界。Reynar[8]提出了适于主题划分的几个特征: 1) 特定领域的提示短语, 例如在广播新闻文本中, joining us, welcome back 等短语都可以预示新的主题; 2) 二元词组频率, 使用二元词组频率可以避免单词频率引发的歧义问题; 3) 命名实体的重复, 同一个命名实体在不同块中出现, 那么这些块极有可能讨论的是同一主题; 4) 代词处理, 可以假定, 当某个句子被视作该主题的第一句, 而其第一个词是代词, 那么很有可能这个句子就不该成为该主题的首句, 即边界识别有误。

融合特定语言现象和文本特征的方法, 利用了更多的文本信息, 是对第一类方法的补充。该方法的主要不足在于, 一些特征具有领域依赖性, 如提示性短语, 在不同领域、不同语种中区别较大, 适应面较窄。

第三类是概率统计模型的使用。这类方法相信合适的概率统计模型能够为片段边界的估计提供可靠依据。随着统计自然语言处理技术的发展, 概率统计模型在自然语言处理的众多领域都开始发挥重要的作用, 主题划分也不例外。

Van Mulbregt 等[13]采用隐马尔科夫模型进行主题划分, Blei 等[14]在其基础上, 引入了一种结合方位模型与隐马尔科夫模型的特殊概率模型来进行主题化分。石晶等[15]分别采用 PLSA、LDA 以及小世界模型来进行主题划分, 并从模型的特点、分割策略以及实验结果等方面对 3 种方法进行了比较, 结果表明基于 LDA 模型的分割比基于 PLSA 模型的分割具有更大的稳定性, 且分割效果更好, 基于小世界模型的分割策略更适合小世界模型特性明显的文本。使用概率统计模型对文本主题进行划分, 关键是选择合适的模型。针对不同文本特征, 如何选择模型是一个难点。使用概率统计模型的优点在于其可移植性较强, 不足之处是计算复杂性较高。

3 面向自动文摘的主题划分方法

3.1 主题的呈现方式

主题(即切分块)的呈现方式直接影响着主题划分方法和效果。Salton 等[16]认为主题片段应该由段落构成, 一个切分块是文本的一个连续段落集, 这个块内段落之间的联系很紧密, 但与其相邻的块联系较少。语言学理论也支持这个观点[12], 从语言学角度来看, 段落标记符通常比句子标记符更能标志一个子主题的完成。Ponte 等[17]则认为主题是几个句子组成的集合, Choi[10]也支持这个观点, 他认为受文本内容、作者写作风格和文本呈现方式的影响, 一个段落中可能包含有多个主题。

实际上, 主题以何种方式呈现与文本本身的特征有重要的关联。对于一个章节划分严谨的文本(如学术论文)而言, 主题可以以章节的方式呈现。对于一个段落数量较多的文本, 主题可以以段落集的方式呈现; 而对于一个没有段落标记的长文本, 或者是段落数较少的短文本, 可以认为主题以句子集的方式呈现。

3.2 面向自动文摘的主题划分流程

图 1 为面向自动文摘任务的主题划分总流程图。

对于中文文档而言, 预处理包括分句、分词和停用词删除。

分句: 句子是文摘最基本的组成单位, 在主题划分过程中, 句子或是直接作为主题切分块的组成部分, 或是以段落(句子集)的形式组成主题切分块。因此分句的结果对主题划分效果有着直接的影响。

分词: 汉语是以字而不是词作为语言的基本构成单位, 而词是最小的能够独立活动的有意义的语言成分。分词是中文信息处理的第一步。本文所采用的分词工具是中国科学院计算技术研究所开发的 ICTCLAS 2011, 该系统是基于多层隐马模型的汉语词法分析系统, 分词的正确率高达 97%以上, 是目前应用最广的汉语分词系统。

停用词删除: 一些虚词、叹词、助动词等, 如“的”、“啊”、“些”, 这些词对主题划分和自动文摘无益, 在自然语言处理领域将其称为“停用词”, 汉语文本分词后, 利用事先制定的停用词表, 将文本中的停用词删除。

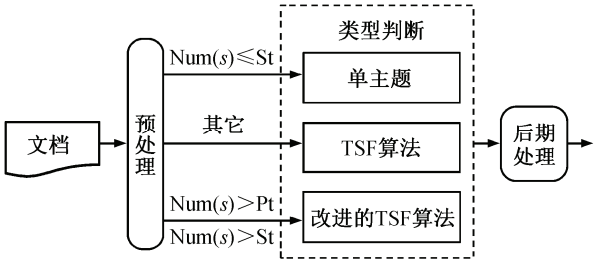

图 1 面向自动文摘的主题划分流程

Fig.1 Flow path of topic partition

类型判断模块主要是根据输入文本的特征对文本采取不同的主题划分策略。对自动文摘而言，我们最终需要从切分的主题中按比例(通常小于 30%)抽取重要的句子，因此主题所包含的句子数不宜过小。我们拟采取的策略是设置两个阈值，句子数 St 和段落数 Pt。如果文本句子数小于 St，则认为该文本讨论的是同一个主题，直接进行抽取；如果文本段落数大于 Pt 且句子数大于 St，则按段落划分主题；否则，认为主题以句子集形式呈现，按句子划分主题。

3.3 主题划分算法

在第 2 节中，我们提到了目前主题划分的几种主流方法。Kern 等[5]指出 TextTiling 算法虽然复杂度较低，但划分效果不佳，而 C99 算法虽然划分效果较好，但计算复杂性较高。他们提出了一种主题划分方法 TSF(TextSegFault)算法，该算法在很多方面与 TextTiling 算法相近，也是一种基于词汇衔接理论的方法。根据文章呈现的评价结果，TSF 算法在切分效果上要远好于 TextTiling 算法，并且只有 $O(n)$的计算复杂度。

TSF 算法默认主题是由句子集组成，本文对 TSF 算法进行扩展，包含句子集主题划分和段落集主题划分两种。句子集主题划分算法描述如下。

1) 在每两个句子之间设置一个位置 i，给定窗口大小 W，W 反应了切分的最小主题所含的句子数。在每个 i 位置，建立 W 大小的前驱块 B_i^{pre} 和后继块 B_i^{post}。

2) 位置 i 的内部块相似度记作 IS_i，定义如下:

$$\text{IS}_i = \frac{\mu(B_i^{\text{pre}}, B_i^{\text{pre}}) + \mu(B_i^{\text{post}}, B_i^{\text{post}})}{2}, \quad (1)$$

其中 μ 表示两个块之间的平均相似度。位置 i 的外部块相似度记作 OS_i，定义如下:

$$\text{OS}_i = \mu(B_i^{\text{pre}}, B_i^{\text{post}})。 \quad (2)$$

3) 计算位置 i 的不相似度

$$\text{disSim}_i = \frac{\text{IS}_i - \text{OS}_i}{\text{IS}_i} \quad (3)$$

设定一个阈值 q，如果 i 点的不相似度大于 q，则将该点视作主题边界的候选点。

4) 为了避免连续的位置出现高的不相似度，我们采用前向查看的方法，查看距离为 W，如果该点的不相似度小于下一个点的不相似度，则此点不作为主题边界点。

段落集主题划分认为每个块由段落组成，在段落与段落之间设置位置点，针对每个位置点，计算其前驱段落和后继段落的内部块平均相似度以及外部平均相似度，然后计算每个点的不相似度，利用前向查看得到主题边界点。采用 TSF 算法得到的主题是由句子集组成，而采用改进的 TSF 算法(段落集主题划分)进行主题划分得到的主题是由段落集组成的。

从算法描述不难看出，算法的核心在于句子相似度计算。最经典的句子相似度计算方法是利用向量空间模型将每个句子表示为向量，通过计算向量之间的余弦，确定句子之间的相似度。计算公式如下:

$$\text{Sim}(S_1, S_2) = \frac{\sum_{i=1}^{n} W_{1i} \cdot W_{2i}}{\sqrt{(\sum_{i=1}^{n} W_{1i}^2)(\sum_{i=1}^{n} W_{2i}^2)}}, \quad (4)$$

其中，n 为整个文档中的词数，W_{1i} 为 S_1 中第 i 个词项的权重，W_{2i} 是 S_2 中第 i 个词项的权重。

关于词项的权重，算法使用的是改进的 TF-IDF 算法，句子 j 中的词项 i 的权重 $W_{i,j}$ 的计算公式如下:

$$W_{i,j} = \frac{\sqrt{\text{tf}_{i,j}}}{\text{len}_j}(\log(\frac{\text{size}}{\text{tf}_i}) + 1), \quad (5)$$

其中，$\text{tf}_{i,j}$ 为词项 i 在句子 j 中出现的次数，len_j 为句子 j 的长度(即含词项的个数)，size 为参考语料中的词汇总数，tf_j 语料中词项 i 出现的次数。

对词项权重进行归一化后得到

$$W'_{i,j} = \frac{W_{i,j}}{\sqrt{\sum_{i=1}^{|\text{len}_j|} {W_{i,j}}^2}}。 \quad (6)$$

对主题划分结果的后期处理主要是利用“代词特征”来合并一些不合理切分。一些代词通常不会是主题中的第一个词汇，如“他”、“这些”等。本文通过语料库查询的方式，构建了一个词典库，如果切分的主题首词出现在词典库中，则认为该切分是一个错误的切分。考虑到自动文摘对主题划分的颗粒度要求不高，可以直接将该切分点前后两个主题进行合并。

4 主题划分结果评价

对主题划分结果进行评价是一件很困难的事，

一篇多主题文章即便是我们人为来判断主题边界, 其结果也通常有所差异, 因此很难找出一个参考的切分结果以供比较。Reynar[11]将多个文本连结到一个文本中, 将主题划分的结果与文本之间的边界做比较来对主题划分算法进行评价。这种方法虽然会有一个准确的参考结果, 但正如我们前文分析的那样, 寻找文本边界和主题划分是两个不同的任务。Labadie 等[1]研究发现, 一些算法对文本边界识别有效, 但对于主题划分则不然。

此外, 面向不同任务的主题划分, 其评价方式也应该有所不同。例如, 对自动文摘而言, 将主题边界划到实际边界之前或之后的几个句子处(此现象称为 near miss), 不会对文摘结果产生太大的影响。但是, 对于新闻边界识别而言, 必须找到准确的边界。NenKova 等[18]等建议, 不要直接对划分结果进行评价, 而是考虑其对任务的影响。

准确率和召回率是信息检索领域两个重要的评价指标。在主题划分领域, 准确率是指总识别的边界数中识别正确的边界数所占的百分比, 召回率是指识别正确的边界数与参考划分的边界数的百分比。然而准确率和召回率在主题划分领域并不是好的评价指标。其一是它们之间存在一定的冲突, 提高准确率在一定程度上会降低召回率, 反之亦然, 虽然 F 值评价法从一定程度上平衡了准确率和召回率, 但其可理解性不强。其二是无法处理 near miss 现象。图 2 显示了 A-0 和 A-1 两种不同的主题划分算法得出的切分结果。如果采用准确率和召回率进行评价, 其准确率和召回率均为 0。事实上, 从图中不难看出, A-0 算法得出的结果要好于 A-1 算法。

针对 near miss 的情况, 本文提出了一种简单的切分结果衡量机制——切分合理度 R。计算方式如下。

如果切分的主题个数 n 小于或等于参考切分主题个数 N, 则

$$R = \sum_{i=1}^{n} P(i) / N \ , \tag{6}$$

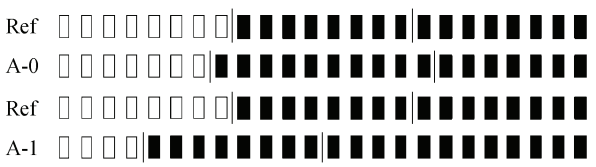

图 2　参考切分和实际切分

Fig. 2　Reference partition and partition by A-0&A-1 Algorithm

如果切分的主题个数 n 大于实际主题个数 N, 则

$$R = N \sum_{i=1}^{n} P(i) / n^2 \ , \tag{7}$$

其中 $P(i)$ 表示每个实际切分点的得分, 记当前切分点离参考切分点的最短距离为 $\min \mathrm{dis}(i)$, 则

$$P(i) = 1/(1 + \min \mathrm{dis}(i)) \tag{8}$$

以上图为例, A-0 算法得到的 R 值为 0.5, A-1 算法得到的 R 值为 0.13。

本文采用的参考语料和测试语料均来自 163 Chinese News Dataset 2011 新闻数据库[19]。该语料库中的每条记录包括新闻标题、内容、日期和核心提示等。我们选取了 30 个 T1 类型的文本(句子数大于 St, 但段落数小于 Pt), 50 个 T2 类型的文本(句子数大于 St 且段落数大于 Pt)进行测试。所选取的这些新闻文本都具有子标题, 根据子标题人工对文本进行切分得到文本的一个参考切分。之后删除文本中的子标题, 利用文章的主题划分方法对其进行主题划分, 将主题划分结果与参考切分进行比较, 利用切分合理度 R 对每个文档的切分结果进行评估, 然后取其平均值 $R' = \sum R / C$, 结果如表 1 所示。

从切分结果来看, 平均召回率较高, 而准确率相对较低, 造成这一现象的主要原因是实际切分数大于参考切分数, 这与参考切分的切分颗粒度相关。这也直接导致了平均切分合理度较低, 然而对自动文摘而言, 这样的切分合理度是有效的, 它保证了文摘的全面性。实验中涉及到的参数设置如表 2 所示。

表 1　实验结果评价

Table 1　Evaluation of the result

类型	主题表征	文本数量 C	平均切分合理度 R'	平均准确率/%	平均召回率/%
T1	句子集	30	0.58	78.6	95.2
T2	段落集	50	0.67	83.8	97.5

表 2　实验中用到的参数

Table 2　Parameters used in experiment

参数名称	Pt	St	阈值 q_1(句子切分)	阈值 q_2(段落切分)	窗口大小 W
参数值	9	8	0.45	0.8	3

5　总结与思考

本文利用改进的 TSF 算法, 面向自动文摘任务

对文档进行了主题划分。利用主题划分结果，从不同主题中抽取句子形成文摘可以提高文摘的全面性。从算法上看，本文所选取的实验参数是通过人为设定的，未考虑参数的变化对实验结果带来的影响。事实上，这一点是非常重要的。除了相关参数，计算词项权重时所用到的参考语料库的大小也会对实验结果产生一定的影响，这些都是下一步可以研究讨论的。

此外，对算法的进一步改进，可以集中在句子相似度计算和词汇权重计算上。同时，可以加入一些语言特征，对划分结果进行修正。就自动文摘而言，对于划分的主题，还需要计算主题权重，以方便计算需要从该主题中抽取的句子数。总之，可以预见，主题划分的引入能提高文摘的质量。

参考文献

[1] Labadié A, Prince V. Finding text boundaries and finding topic boundaries: two different tasks // Nordström B, Ranta A. GoTAL 2008. Gothenburg, Sweden, 2008, 5221: 260–271

[2] 傅间莲，陈群秀. 自动文摘系统中的主题划分问题研究. 中文信息学报, 2005, 19(6): 25–28

[3] Brown G, Yule G. Discourse analysis, Cambridge Textbooks in Linguistics Series. Britain: Cambridge University Press, 1983

[4] Reynar J. Statistical models for topic segmentation // Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. New Jersey, USA,1999: 357–364

[5] Kern R, Granitzer M. Efficient linear text segmentation based on information retrieval techniques // MEDES. Lyon, France, 2009: 167–171

[6] Hearst M. TextTiling: segmenting text into multi-paragraph subtopic passages. Computational Linguistics, 1997, 23(1): 33–64

[7] Beeferman D, Berger A, Lafferty J. Statistical models for text segmentation. Machine Learning, 1999, 34:177–210

[8] Halliday M A K, Hasan R. Cohesion in English. London: Longman, 1976

[9] Hearst M A, Plaunt C. Subtopic structuring for full-length document access // Proceedings of the 16th Annual International ACM/SIGIR. Pittsburgh, PA, 1993: 59–68

[10] Choi F Y Y. Advances in domain independent linear text segmentation // NAACL. Seattle, Washington, 2000: 26–33

[11] Reynar J C. An automatic method of finding topic boundaries // Proceedings of the 32nd Annual Meeting (student session). Las Cruces, NM, 1994: 331–333

[12] Chafe, Wallace L. The flow of thought and the flow of language // Givon T. Syntax and Semantics: Discourse and Syntax. New York: Academic Press, 1979: 59–182

[13] Van Mulbregt P, Carp I, Gillick L, et al. Text segmentation and topic tracking on broadcast news via a hidden markov model approach // Proceedings ICSLP-98.Sydney, Australia, 1998: 2519–2522

[14] Blei D, Moreno P. Topic segmentation with an aspect hidden Markov model // Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Louisiana, USA: ACM Press, 2001:343–348

[15] 石晶，李万龙. 三种主题分割算法的对比研究. 计算机工程与应用, 2009, 45(18): 135–151

[16] Salton G, Allan J, Buckley C, et al. Automatic analysis, theme generation, and summarization of machine-readable texts. Science, 1994, 264: 1421–1426

[17] Ponte J M, Croft W B. Text segmentation by topic // European Conference on Digital Libraries. Pisa, Italy, 1997: 113–125

[18] NenKova A, McKeown K. Automatic summarization. Fundations and Trends in Information Retrieval, 2011, 5: 103–233

[19] Liu Zhiyuan, Chen Xinxiong, Zheng Yabin, et al. Automatic keyphrase extraction by bridging vocabulary gap // Proceedings of CoNLL. Portland, Oregon, USA, 2011: 135–144