

# 面向网络演化信息的动态文摘方法研究

张 瑾<sup>1,2)</sup> 许洪波<sup>1)</sup> 程学旗<sup>1)</sup>

<sup>1)</sup>(中国科学院计算技术研究所 北京 100080)

<sup>2)</sup>(中国科学院研究生院 北京 100039)

**摘 要** 随着互联网的发展和 Web2.0 的出现,网络信息内容的动态演化性越来越明显.该文从网络信息的时间演化性出发,给出了动态文摘的形式化定义.在分析当前信息与历史信息的演化关系的基础上,采用内容过滤的方法度量演化内容的差异性,从而得到三种动态文摘模型,并基于模糊隶属度给出了具体的动态文摘生成方法.在 DUC 2007 测试数据上的实验,证明了文中所提出动态文摘模型及生成方法的有效性.

**关键词** 动态演化性;动态文摘;模型;性能

**中图法分类号** TP391

## Research on Dynamic Summarization for Evolutionary Web Information

ZHANG Jin<sup>1,2)</sup> XU Hong-Bo<sup>1)</sup> CHENG Xue-Qi<sup>1)</sup>

<sup>1)</sup>(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)

<sup>2)</sup>(Graduate University of Chinese Academy of Sciences, Beijing 100039)

**Abstract** With the development of Internet and the emergence of Web2.0, the dynamic evolution of Web information is becoming more and more evident. This paper presents the formalization of dynamic summarization according to temporal evolution of Web information. Based on the analysis of evolution of Web information, the authors measure the difference between current information and history information, and then propose three models for dynamic summarization and the solutions in detail. Experiments on DUC 2007 update dataset illustrate the performance of the methods.

**Keywords** dynamic evolution; dynamic summarization; model; performance

## 1 引 言

多文档自动文摘技术作为一种提炼概要信息的有效手段,已经得到了广泛的研究.传统的多文档文摘技术是一种静态文摘,即针对某个封闭的静态文档集生成摘要,不考虑文档集的对外联系.在 Web2.0 时代,出现在 bbs 论坛、blog、在线评论等新媒体中的网络信息(如网络话题、热点事件等,表现

为一系列相关文章的集合)是动态演化的,它们随着时间的变化而出现、发展直至消亡,一个话题在不同的时刻具有不同的侧重点,而不同时刻的话题内容之间具有关联性.因此,如何对动态演化的网络信息进行文摘成为一个新的研究课题.

我们将针对网络演化信息的水摘称为动态文摘,它是一种具有时序偏向的多文档文摘,其研究对象是网络动态演化信息的关联文档集.如果把动态演化信息的生存时间  $T$  划分为  $n$  个时间段  $t_1$ ,

$t_2, \dots, t_n$ , 各个时段包含的文档集分别为  $D_1, D_2, \dots, D_n$ , 则动态文摘问题可以形式化为: 已知  $t_1, \dots, t_{i-1}$  时段的文档集  $D_1, \dots, D_{i-1}$ , 求  $t_i$  时段的文档集  $D_i$  的摘要, 即  $DynSummary(D_i | D_1 \dots D_{i-1}), 1 \leq i \leq n$ , 如图 1 所示. 当  $i=1$  时, 该问题退化为传统的静态文摘问题.

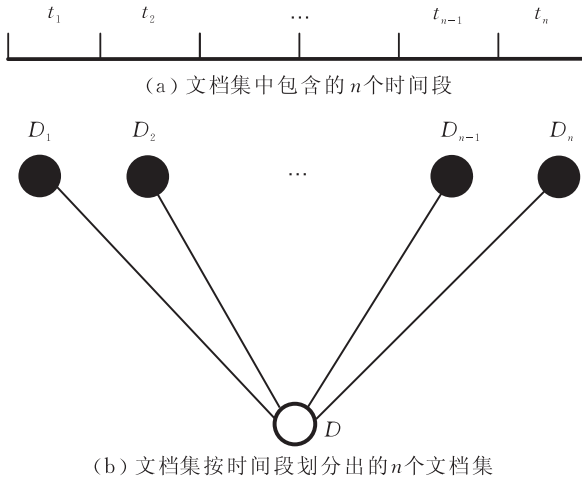


图 1 动态文摘的形式化表示

动态文摘是传统静态文摘的延伸和扩展, 除了需要保证文摘信息的主题相关性和内容的低冗余性之外, 还需要针对内容的动态演化性分析已出现信息和新出现信息的关系, 消除旧信息, 摘要新信息, 使文摘随话题的演化而动态更新. 根据动态文摘问题的形式化描述, 本文从分析历史信息与当前信息关系的角度提出了三种动态文摘基本模型, 并在此基础上, 利用我们提出的静态文摘算法 GSPSummary, 给出了可行的动态文摘生成方法.

本文第 2 节对动态文摘的相关工作进行了介绍; 第 3 节针对内容的动态演化性提出了三种动态文摘模型; 第 4 节在静态文摘方法的基础上给出了动态文摘的生成方法; 第 5 节在 DUC 2007 数据集上对动态文摘方法的性能进行了测试, 并通过与 DUC 2007 Update 任务<sup>①</sup>国际评测结果的对比, 验证了方法的有效性; 最后给出了本文的结论与展望.

## 2 相关工作

动态文摘与静态文摘方法的最大区别在于动态文摘需要在主题相关性的基础上考虑多个文档集之间的时序关系, 分析已出现信息和新出现信息的关系, 从而对内容的动态演化性进行建模和动态文摘的生成.

动态内容的时序划分是动态文摘的基础, 相关研究在新闻事件检测 (News Information Detec-

tion, NID)<sup>[1]</sup> 和 TDT<sup>②</sup> 等领域<sup>[2-4]</sup> 得到了较多关注, Mani 等人使用时域分析方法对新闻事件的内容进行分析<sup>[5]</sup>, Allan 等人借用图形学领域的时间线的构建来进行内容划分<sup>[6]</sup>.

Allan 等人<sup>[7]</sup> 在 TDT 研究的基础上, 探讨了基于内容有用性 (useful) 与新颖性 (novel) 的时域文摘研究方法, 其提出的时序文摘不同于本文的动态文摘, 本质上是一种基于句子排列策略改进的静态文摘.

DUC 2007 国际评测的先导任务 Update Task 实际上就是一个动态文摘问题. 这一任务主要来源于信息检索系统、问答系统和文摘系统中对用户行为的模拟. 该任务假定用户对某个场景已出现的内容有了足够的了解, 在后续文摘中重点关注那些新出现的内容. 因此, 当得到与此场景相关的新信息时, 需要分析新信息与旧信息的关系并生成更新文摘 (Update Summary).

## 3 动态文摘的建模

动态文摘的关键问题是如何对动态信息的演化性内容进行表示, 具体来说就是对时序文档集中当前文档集  $D_i$  与历史文档集  $D_1, \dots, D_{i-1} (1 \leq i \leq n)$  之间的内容差异性如何建模. 为了方便叙述, 首先给出如下定义.

**定义 1.** 把时序文档序列中当前文档集  $D_i$  包含的信息称为当前信息 (current information), 用  $I_c$  表示.

**定义 2.** 把时序文档序列中历史文档集  $D_1, \dots, D_{i-1} (1 \leq i \leq n)$  包含的信息称为历史信息 (history information), 用  $I_h$  表示.

**定义 3.** 用  $f$  表示从文档空间到文摘空间的映射关系, 则时序文档序列中任一文档集  $D_i$  的文摘可记为  $f(D_i)$ , 同样的, 历史信息  $I_h$  的文摘称为历史文摘, 表示为  $f(I_h)$ , 当前信息  $I_c$  的文摘称为当前文摘, 表示为  $f(I_c)$ .

在以上定义的基础上, 可以把动态文摘问题  $DynSummary(D_i | D_1 \dots D_{i-1}), 1 \leq i \leq n$  转化为对历史信息  $I_h$  和当前信息  $I_c$  之间演化内容的差异性建模和求解. 我们对历史信息  $I_h$  和当前信息  $I_c$  的演变关系进行了分析, 采用内容过滤的方法刻画演化内容的差异性. 根据被过滤对象的不同, 提出以下三种

① <http://www-nlpir.nist.gov/projects/duc/duc2007/tasks.html#pilot>

② <http://www.nist.gov/speech/tests/tdt/>

基于内容过滤的动态文摘模型.

3.1 文档过滤模型 DFM

动态文摘的基本要求是不重复历史信息,因此,最直接的方法是先找出当前信息中的新信息,然后采用静态文摘方法对新信息生成文摘.新信息可以通过从当前信息  $I_c$  中过滤掉与历史信息  $I_h$  相重叠的内容得到,表示为  $I_c - I_h$ ,然后利用静态文摘方法生成动态文摘  $f(I_c - I_h)$ .这种动态文摘模型从文档内容过滤的角度提取动态信息以生成文摘,我们将其称为文档过滤模型(Document Filtering Model, DFM).以历史信息  $I_h$  为过滤对象的动态文摘模型记为 DFM1.考虑到文摘对文档内容的代表性,为了节省计算代价,可以将过滤对象  $I_h$  替换为历史文摘  $f(I_h)$ ,从而得到一种类似的动态文摘模型  $f(I_c - f(I_h))$ ,记为 DFM2.

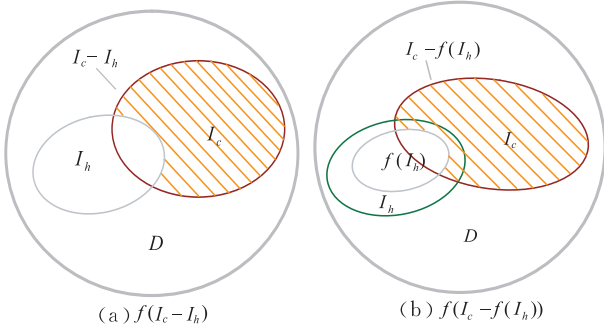


图 2 文档过滤模型

3.2 文摘过滤模型 SFM

第二种内容过滤的思路是首先利用静态文摘方法对当前信息  $I_c$  生成候选文摘  $f(I_c)$ ,然后再从候选文摘中过滤掉与历史信息  $I_h$  的重叠内容,从而得到所需的动态文摘  $f(I_c) - I_h$ .由于这种动态文摘模型以当前文摘  $f(I_c)$  为被过滤对象,我们将其称为文摘过滤模型(Summary Filtering Model, SFM).同样的,以历史信息为  $I_h$  为过滤对象的动态文摘模型记为 SFM1.而以历史文摘  $f(I_h)$  为过滤对象的动态文摘模型  $f(I_c) - f(I_h)$ ,记为 SFM2.

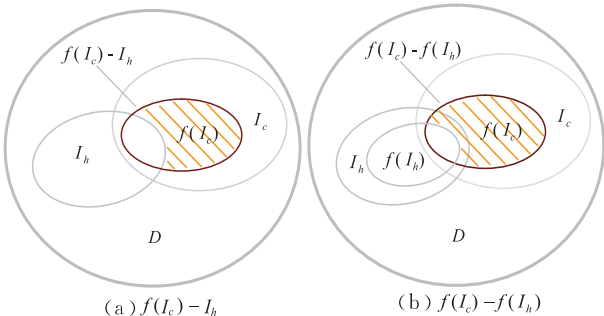


图 3 文摘过滤模型

3.3 合并过滤模型 UFM

由于历史信息和当前信息都是关联于同一主题的,二者虽然时序不同,但内容上存在着一定的关联性.无论是文档过滤模型 DFM 还是文摘过滤模型 SFM,都重点强调了当前信息与历史信息的差异性,而没有充分利用二者之间的关联性.考虑到当前信息与历史信息之间的关联关系,可以得到第三种动态文摘模型:首先对历史信息  $I_h$  和当前信息  $I_c$  合并的全文档空间生成文摘  $f(I_h + I_c)$ ,再从中进行历史信息的过滤,从而生成动态文摘.我们将这种合并两类信息再进行内容过滤的动态文摘模型称为合并过滤模型(Union Filtering Model, UFM).其中,以历史信息为  $I_h$  为过滤对象的动态文摘模型  $f(I_h + I_c) - I_h$  记为 UFM1,而以历史文摘  $f(I_h)$  为过滤对象的动态文摘模型  $f(I_h + I_c) - f(I_h)$ ,记为 UFM2.

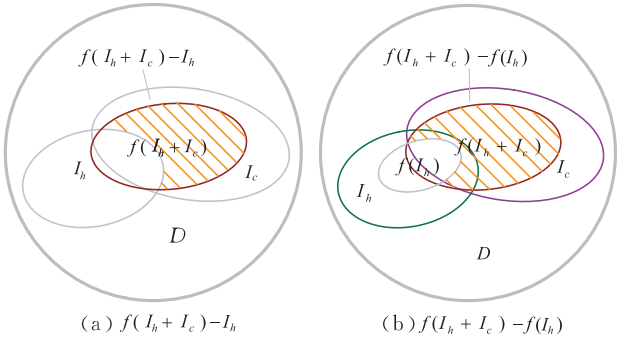


图 4 合并过滤模型

4 动态文摘的生成方法

4.1 动态文摘生成的思路

本文讨论的文摘是基于句子的抽取式文摘.我们用  $S_{d_j}$  表示文档  $d_j$  中所有句子的集合,用  $S_{D_i}$  表示文档集  $D_i$  中所有文档包含的句子的集合,即

$$\begin{cases} S_{d_j} = \{s | s \in d_j\} \\ S_{D_i} = \bigcup S_{d_j} = \{s | s \in d_j, d_j \in D_i\} \end{cases}.$$

于是,当前信息  $I_c$  可以表示为当前文档集  $D_i$  所包含的句子的集合,而历史信息  $I_h$  则可以表示为所有已出现文档集  $D_1, \dots, D_{i-1}$  中所包含的句子的集合,即

$$\begin{cases} I_c = S_{D_i} \\ I_h = \bigcup_{j=1}^{i-1} S_{D_j} \end{cases}.$$

在前面提出的动态文摘过滤模型中,关键是如何度量当前信息和历史信息的差异性,即如何计算  $I_c - I_h$ .我们借用模糊数学中的隶属度(Degree)概

念引入了一个新的运算符 $\tilde{\cap}$ ,称为模糊与运算,其计算方法如下:

$I_c \tilde{\cap} I_h = \{s | \text{similarity}(s, s_k) \geq \text{Degree}, s \in I_c, s_k \in I_h\}$ ,  
即  $I_c$  和  $I_h$  的模糊与集合由那些相似度超过某一隶属度的句子构成. 其中  $\text{Degree} \in [0, 1]$  是隶属度, 当  $\text{Degree} = 1.0$  时,  $\tilde{\cap}$  运算的结果是  $I_c$  和  $I_h$  的交集, 即  $I_c \tilde{\cap} I_h = I_c \cap I_h$ ; 当  $\text{Degree} = 0.0$  时,  $\tilde{\cap}$  运算的结果是  $I_c$ , 即  $I_c \tilde{\cap} I_h = I_c$ .

因而,可以得到  $I_c - I_h$  的计算方法为

$$I_c - I_h = \{s | s \in I_c, s \notin I_c \tilde{\cap} I_h\}.$$

文档空间到文摘空间的映射关系  $f$  实际就是一个静态文摘算法,可以是任何已有的多文档文摘算法,本文中采用我们自主提出的基于子主题图划分模型的多文档文摘方法 GSPSummary,关于该算法的细节参见文献[8]. 则  $f(I)$  可以表示为

$$f(I) = \text{GSPSummary}(I).$$

具体地,针对前文提出的三种过滤模型,可以分别给出动态文摘生成方法.

#### 4.2 基于文档过滤的动态文摘生成方法

文档过滤模型首先要计算当前信息  $I_c$  和历史信息  $I_h$  或者历史文摘  $f(I_h)$  的差集,然后利用 GSPSummary 算法生成差集的文摘作为动态文摘. 其两种实现模型 DFM1 和 DFM2 的动态文摘计算公式分别为

$$\begin{aligned} f(I_c - I_h) &= \text{GSPSummary}(I_c - I_h) \\ &= \text{GSPSummary}\{s | s \in I_c, s \notin I_c \tilde{\cap} I_h\} \end{aligned} \quad (1)$$

$$\begin{aligned} f(I_c - f(I_h)) &= \\ &= \text{GSPSummary}(I_c - \text{GSPSummary}(I_h)) \end{aligned} \quad (2)$$

#### 4.3 基于文摘过滤的动态文摘生成方法

文摘过滤模型先生成当前信息  $I_c$  的文摘,然后根据历史信息  $I_h$  或者历史文摘  $f(I_h)$  对其进行过滤. 其两种实现模型 SFM1 和 SFM2 的动态文摘计算公式分别为

$$f(I_c) - I_h = \text{GSPSummary}(I_c) - I_h \quad (3)$$

$$\begin{aligned} f(I_c) - f(I_h) &= \text{GSPSummary}(I_c) - \\ &= \text{GSPSummary}(I_h) \end{aligned} \quad (4)$$

#### 4.4 基于合并过滤的动态文摘生成方法

合并过滤模型首先计算当前信息  $I_c$  和历史信息  $I_h$  的合集:

$$I_c + I_h = \{s | s \in I_c \cup I_h\}.$$

然后生成合集的文摘,最后根据历史信息  $I_h$  或

者历史文摘  $f(I_h)$  对其进行过滤. 其两种实现模型 UFM1 和 UFM2 的动态文摘计算公式分别为

$$f(I_c + I_h) - I_h = \text{GSPSummary}(I_c + I_h) - I_h \quad (5)$$

$$\begin{aligned} f(I_c + I_h) - f(I_h) &= \text{GSPSummary}(I_c + I_h) - \\ &= \text{GSPSummary}(I_h) \end{aligned} \quad (6)$$

## 5 实验结果与分析

我们在国际评测 DUC 2007 Update 任务的公开数据集上进行了实验测试,以验证动态文摘方法的有效性. DUC 2007 Update 文摘语料来源于 TREC 评测中的 AQUAINT 语料,其中包括 10 个主题,每个主题包含 25 篇文档,这 25 篇文档又被按照事件的发展分成了 3 个子集 A, B, C. 每个主题中 3 个子集内文档所描述的内容,存在以下关系: A 所描述事件的状况先于 B 和 C, B 所描述事件的状况先于 C, 即 A 是 B 和 C 的历史信息, B 是 C 的历史信息. 动态文摘的目标就是在给定历史信息的前提下分别对 A, B, C 生成文摘,其中 A 的历史信息为空. 评价标准采用文摘评测领域著名的 ROUGE 工具<sup>[9]</sup>,其中最主要的两个评价指标是 ROUGE-2 和 ROUGE-SU4. 由于 DUC 2007 Update 任务实际上也是针对动态文摘的评测,所以我们将动态文摘结果的 ROUGE-2(R-2)和 ROUGE-SU4(R-SU4)得分与 DUC 2007 Update 实际系统的得分进行了对比,结果表明我们的动态文摘方法具有良好的性能.

表 1 是在不同的隶属度下三种模型生成的动态文摘的 ROUGE 评价结果. 从表中可以看出,当采用精确过滤( $\text{Degree}$  为 1.0)获取差异性内容时,DFM1 文档过滤模型的文摘性能最好;而在模糊过滤时(取隶属度  $\text{Degree} = 0.8$ ),UFM1 合并过滤模型取得最好的性能. 通过三种模型的 ROUGE 得分的对比,可以看出 DFM1 和 DFM2, SFM1 和 SFM2 的文摘性能差别不大,说明文档过滤模型和文摘过滤模型对隶属度的变化不是非常敏感. 由于 DFM2、SFM2 模型的计算代价较小,因此可以作为 DFM1、SFM1 模型的有效代替. 合并过滤模型对隶属度的变化较为敏感,而 UFM1 模型受隶属度的影响又大于 UFM2 模型. 从上面结果可以看出,因为合并过滤模型综合考虑了当前信息和历史信息的相互关系,因此它能够较为有效地进行动态内容的过滤,从而获得较好的文摘质量.

表 1 三种模型的文摘性能对比

Degree	DFM1		DFM2		SFM1		SFM2		UFM1		UFM2	
	R-2	R-SU4	R-2	R-SU4	R-2	R-SU4	R-2	R-SU4	R-2	R-SU4	R-2	R-SU4
1.00	<b>0.1058</b>	<b>0.1456</b>	0.1056	0.1455	<b>0.1058</b>	<b>0.1456</b>	0.1056	0.1455	0.0855	0.1241	0.0797	0.1196
0.80	0.1061	0.1454	0.1056	0.1455	0.1057	0.1459	0.1056	0.1455	<b>0.1078</b>	<b>0.1464</b>	0.0811	0.1216

表 2 是隶属度分别取 1.0 和 0.8 时三种模型的最好结果与参加 DUC 2007 Update 任务评测的前三名实际系统的性能对比. 其中 DFM1/SFM1 ( $Degree=1.0$ ) 和 UFM1 ( $Degree=0.8$ ) 在 ROUGE-2 指标上的得分比排名第一的 LCC 系统稍差, 但远优于排名第二的 IIIT 系统; 而在 ROUGE-SU4 指标上, 三种模型的最好性能均超过 LCC 系统, 可以排在 DUC 2007 评测的第一名, 这说明我们的动态文摘方法具有很好的性能.

表 2 与 DUC 2007 实际系统的性能对比

系统	R-2	R-SU4
DFM1/SFM1 ( $Degree=1.0$ )	<b>0.1058</b>	<b>0.1456</b>
UFM1 ( $Degree=0.8$ )	<b>0.1078</b>	<b>0.1464</b>
LCC(Rank 1)	<b>0.1119</b>	<b>0.1431</b>
IIIT(Rank 2)	0.0985	0.1352
NUS(Rank 3)	0.0962	0.1325

为了研究隶属度对动态文摘的影响, 我们对受隶属度影响最大的 UFM1 模型进行了分析. 实验在 0.3~1.0 的取值范围内以 0.1 的步长对隶属度进行调整, 并用 ROUGE 指标对 UFM1 模型生成的动态文摘进行评价, 实验结果如图 5 所示.

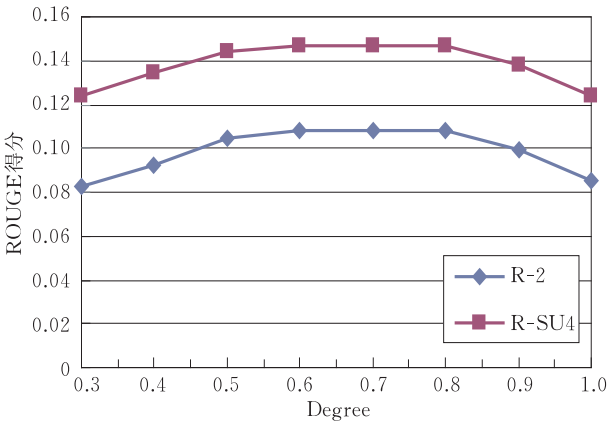


图 5 UFM1 模型受隶属度影响的分析

从图 5 可以看出, UFM1 模型随着隶属度的变化呈现先上升后下降的趋势, 在隶属度为 0.6 时达到最优性能. 此时, 在 ROUGE-2 和 ROUGE-SU4 上的得分分别为 0.1081 和 0.1472, 在 ROUGE-2 上进一步缩小了与 LCC 的差距, 同时在 ROUGE-SU4 上更加凸显了 UFM1 模型的优越性. 通过以上实验结果, 可以说明基于合并过滤的模型确实能够

较为有效地过滤冗余的历史信息, 获取有用的动态信息生成动态文摘, 从而获得较好的动态文摘质量. 因此, 基于隶属度的合并过滤模型生成文摘的方法是一种较为有效的动态文摘方法.

6 结 论

面向网络演化信息的动态文摘是一个全新的研究课题, 目前正处于起步阶段. 我们在参加 DUC 2007 Update 国际评测的基础上, 对动态内容的演化关系进行了分析, 采用内容过滤的方法刻画演化内容的差异性, 从而提出了三种动态文摘模型, 并给出了基于模糊隶属度的动态文摘生成方法. 在 DUC 2007 测试数据上进行的实验证明了我们所提出的动态文摘模型及生成方法的有效性. 我们的下一步工作是, 进一步提高静态文摘算法的性能, 分析不同的动态内容时域分析方法, 将演化信息的内容差异性和主题相关性进行更好的结合等.

参 考 文 献

[1] Allan J, Jin H, Rajman M, Wayne C, Gildea D, Lavrenko V, Hoberman R, Caputo D. Topic-based novelty detection. Center for Language and Speech Processing, Johns Hopkins University, Baltimore; Technical Report ws99, 1999

[2] Allan J, Papka R, Lavrenko V. On-line new event detection and tracking//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, 1998; 37-45

[3] Cao Bin, Shen Dou, Sun Jian-Tao, Wang Xuan-Hui, Yang Qiang, Chen Zheng. Latent factor detection and tracking with online nonnegative matrix factorization//Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07). Hyderabad, India, 2007; 2166-2171

[4] Mei Qiao-Zhu, Zhai Cheng-Xiang. Discovering evolutionary theme patterns from text: An exploration of temporal text mining//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago, Illinois, USA, 2005; 198-207

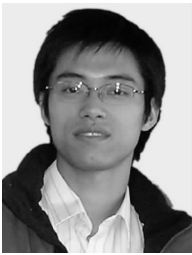
[5] Mani I, Wilson G. Robust temporal processing of news//Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Hong Kong, China, 2000; 69-76

[6] Swan R, Allan J. Automatic generation of overview time-lines//Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Athens, Greece, 2000: 49-56

[7] Allan J, Gupta Rahul, Khandelwal Vikas. Temporal summaries of news topics//Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, Louisiana, United States, 2001: 10-18

[8] Zhang Jin, Xu Hong-Bo, Cheng Xue-Qi. GSPSummary: A graph-based sub-topic partition algorithm for summarization//Proceedings of the 4th Asia Information Retrieval Symposium (AIRS2008). Harbin, China, 2008: 327-341

[9] Lin C-Y. ROUGE: A package for automatic evaluation of summaries//Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004). Barcelona, Spain, 2004



**ZHANG Jin**, born in 1978, Ph. D. candidate. His current research interests include text mining and multi-document summarization.

**XU Hong-Bo**, born in 1975, Ph. D. , associate professor. His current research interests include text mining, Web retrieval and information filtering, etc.

**CHENG Xue-Qi**, born in 1971, Ph. D. , professor. His current research interests include networking and information security, large-scale text content computing and information grid.

Background

The authors are members of Web mining and search group of Institute of Computing Technology, Chinese Academy of Sciences. The major research interests of this group are shallow natural language processing, text mining and information retrieval. The work is supported by the National Basic Research Program(973 Program) of China "Large-Scale Text Content Computing" under grand No. 2004CB318109. With the development of Internet and the emergence of Web2.0,

dynamic summarization becomes an important research question especially with voluminous coverage of a long-running story. This paper presents the definition of dynamic summarization according to temporal analysis and then proposes the fundamental content filtering models for the identification of dynamic information. Furthermore, the solutions for the content filtering models are proposed.