

基于信息融合的多文档自动文摘技术

徐永东 徐志明 王晓龙

(哈尔滨工业大学计算机学院智能技术与自然语言处理研究室 哈尔滨 150001)

摘 要 提出了一个面向多文档自动文摘任务的多文本框架(Multiple Document Framework, MDF),该框架通过系统地描述不同层面的文本单元之间的相互关系以及文档集合蕴含的事件在时间上的发生及演变,将多篇文档在不损失文档集合原有信息的前提下实现信息融合. MDF 简化了传统交叉文本结构理论的文本集合表示模型,又补充了信息融合理论中缺乏的事件主题的演变性和分布性信息. 文中给出了建立 MDF、基于 MDF 的信息融合、文摘生成等一整套算法. 通过对 32 组不同主题的网络文档试验结果表明, MDF 策略很好地实现了多知识源的并行融合,并获得了较好的结果.

关键词 多文本框架;多文档自动文摘;信息融合;时间

中图法分类号 TP391

Multi-Document Automatic Summarization Technique Based on Information Fusion

XU Yong-Dong XU Zhi-Ming WANG Xiao-Long

(Intelligent Technology & Natural Language Processing Laboratory, School of Computer Science and Technology,
Harbin Institute of Technology, Harbin 150001)

Abstract A Multiple Documents Framework (MDF) is proposed for multi-document automatic summarization task. By representing interrelationship between text units at different levels of granularity and the happen and change of various events at time dimension, this framework can achieve information fusion of multi-document while reserve original information of set of related documents. MDF simplifies traditional multi-document representation in cross structure theory and simultaneously, supplements change and distribution informations of events topics which cannot be obtained in information fusion theory. Concretely, a series of algorithms including building MDF, multi-document information fusion based MDF and summarization generation are proposed. The capability of concurrently fusing multiple knowledge sources of MDF strategies is testified by experiments in 32 different sets of net documents and shows good results.

Keywords multiple document framework; multi-document automatic summarization; information fusion; time

1 引 言

自动文摘技术的发展历史可以追溯到 20 世纪 50 年代,从 Lu Hu 1952 年开始自动文摘技术的研

究以来直到现在,人们对它的研究主要集中在对单一文档文本的信息压缩及抽取的研究. 随着近年来网络信息的爆炸式的膨胀,对单篇文档进行摘要处理已经不能满足人们的工作需求. 在此背景下人们提出了结合多文档处理和自动文摘进行信息抽取及

知识挖掘的思路。

多文档自动文摘技术的提出是继单文档自动文摘之后对文本压缩技术的又一挑战,并在近几年随着 DUC 等国际评测会议的连续举办有了较大突破.2001 年 Radev 首先提出了质心的概念^[1],他认为一篇多文档摘要应反映的是一个文档集的中心主题而不是每篇文档的主题,因此文摘的生成应从识别质心开始.如何将这种以篇章为单位的文本信息进行内容重组,并识别其中心主题成为了越来越多的多文档自动文摘技术研究的新热点,其中文本片段聚类技术是被信息融合式多文档自动文摘广泛采用的方法,包括基于段落的聚类^[2-3]、基于句子的聚类^[4-5]等等.其思路基于这样一个基本假设:由于文章描述同类事件,那么在不同文章中重复出现的相似信息可以认为是文章集的重要内容或者重要主题.这种相似信息通常采用文本片段之间的相似关系来确定,然而事实上除了相似关系,文本单元之间还存在着时序关系和修辞关系,因此在信息融合的同时需要充分考虑时间因素在文本集主题识别中的影响以及细化文本单元间的具体关系类别. Radev^[6]在描述其交叉文本结构理论 CST 时提出了 2 个基本数据结构:多文档立方体和多文档图,分别解决了上述问题,2002 年 Zhang^[7-8]分别给出了基于多文档图的文本单元关系识别以及文摘提取的算法.

受交叉文本结构理论的启发,本文提出了一个用于多文本结构分析式文摘的多文本结构 MDF,并在该结构的基础上进行候选文摘句的抽取、文摘句排序及文摘生成等一系列工作. MDF 的贡献在于:

(1)继承了 CST 的特点,用文章内部修辞关系

和文章间的语义关系连接各级节点(文本单元),并加入了时间信息来协助进行文本集主题识别、主题排序,从而为后续工作提供了一个良好的框架.

(2)简化了 CST 的两个基本数据结构,并用一个三维结构完全实现.事实上,CST 的结构 1 描述了文本单元之间的时序关系;结构 2 描述了不同层面的文本单元之间的逻辑关系.这两个结构需要互相补充,互相支撑,才能完整地表达出文档集合的真实面貌.因此和 CST 的两个数据结构相比,MDF 可以集成更多的信息并使得后续工作变得容易、准确.

(3)令人遗憾的是,Zhang 提出的基于 CST 的自动文摘器最后只作为文摘系统 MEAD 的后处理手段.笔者认为,多文档自动文摘是一个系统工程,需要结合多个 NLP 处理技术来完成,如何组织这些技术来完成每个子功能是一个有挑战的课题.串行处理各个子模块是最普遍采用的方法,这种方法的缺陷在于每个子模块的精度误差将全部累计起来影响最终的结果. MDF 不存在这样的问题,在 MDF 的构建阶段,并行地给出了由多个子模块获得的信息.在文摘抽取阶段,本文在 MDF 基础上应用一个全新的基于信息融合的节点加权及选择算法实现文摘句的抽取.这样,一切具有歧义的判断都在最后一个步骤确定,避免了若干中间环节带来的误差累积,为文摘的质量提供了良好的保障.

2 多文本结构 MDF

MDF 是一个多文本表示模型,一个三维立体图状的交叉文本框架(参见图 1).它由代表语言单元

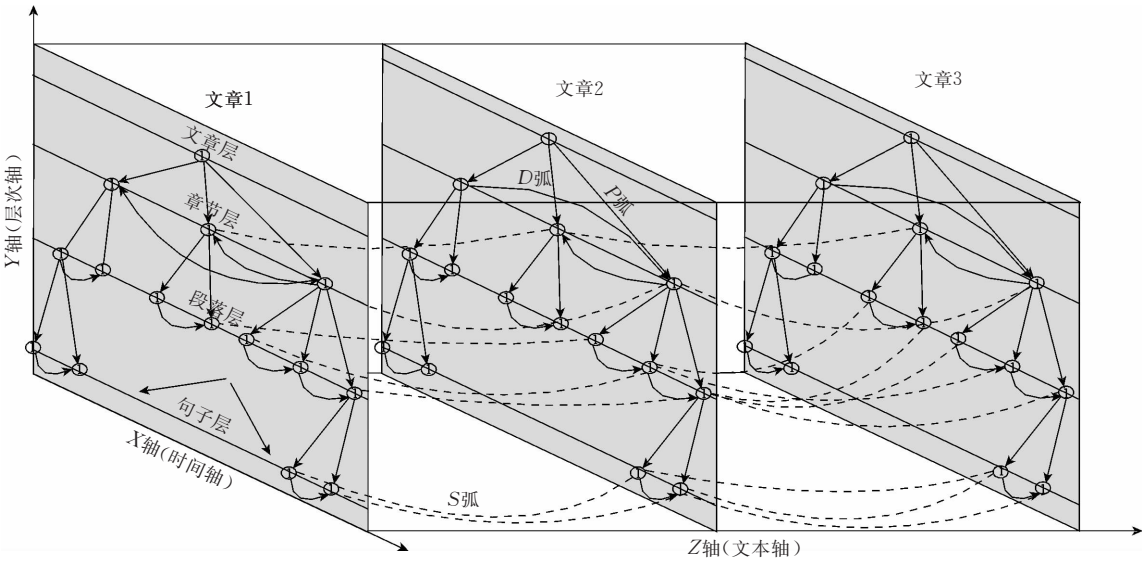


图 1 MDF 三维总体图(图中省略了部分连接弧)

的节点和代表语言单元之间相互关系的连接弧按照特定的方式结合而成,一方面可以方便地不同层面上(包括词、短语、句子、段落、文档)同时表达文本;同时另一方面可以描述多文本内容在时间轴的发展演变。

定义 1. MDF 是一个三维立体图状的结构,可以用一个二元组来表示, $MDF = (V, E)$, V 是节点集, E 是连接不同节点的关系弧集合。

2.1 节点集合 V

MDF 中的节点代表文本集中的各级语言单元,节点全集为 V 。

定义 2. 若 $v \in V$, v 可以用一个三元组来唯一确定, $v = (v_x, v_y, v_z)$, v_x, v_y, v_z 分别表示 v 在 MDF 中的位置坐标。 v_x 表示节点在 x 轴(时间轴)上的位置, v_z 表示 v 所在的文章在 MDF 中的位置, v_y 表示 v 在文章内部的层次, $v_y = \{0, 1, 2, 3\}$, 依次表示全文节点、章节节点、段落节点和句子节点。

2.2 连接弧集合 E

MDF 中的关系弧全集为 E , 包含 3 个子集。 $E = E_p \cup E_d \cup E_s$, 其中, E_p 为部整关系弧集合, E_d 为依存关系弧集合, E_s 为语义关系弧集合。

定义 3. 部整关系弧 P : 给定节点 u, v , 若 $u_z = v_z$, 且 $u_y = v_y - 1$, v 所代表的语言单元是 u 所代表的语言单元的一部分, 则从 u 出发引一条弧指向 v , 并称弧 (u, v) 为 P 弧。 P 弧的全集记为 E_p 。该集合中的元素代表了篇章中相邻上下层语言单元之间的部分-整体关系, 因此本文称之为部整关系弧。

定义 4. 若 $(u, v) \in E_p$, 则称 u 是 v 的上级节点, v 是 u 的下级节点。显然, u 的下级节点集可定义为 $V_u = \{v | (u, v) \in E_p\}$ 。

隶属于同一上级节点的各个下级节点之间存在依存关系, 例如章节内部段落之间的总分关系、段落内部句子之间的因果关系等。在 MDF 中采用依存关系弧(简称 D 弧)来表示依存关系。

定义 5. 依存关系弧 D : 给定节点 u, v, k , 若 $u_z = v_z$, $u_y = v_y$, 且 $u \in V_k, v \in V_k$, 则从 u 出发引一条弧指向 v , 并称弧 (u, v) 为 D 弧。

D 弧从依存关系中起支配作用的语言单元(如总说段、果句等)出发, 指向从属语言单元(如分述段、因句等)。 D 弧的全集计为 E_D 。

不同篇章中隶属于同一级的节点之间存在语义相似关系, 即描述同一事件。在 MDF 中采用语义关系弧(简称 S 弧)来表示语义关系, S 弧没有方向, 连接的两个语言单元之间地位平等。

定义 6. 语义弧 S : 给定节点 u, v , 若 $u_z \neq v_z$,

$u_y = v_y$, 则引一条弧将 u, v 连在一起, 并称弧 (u, v) 为 S 弧。 S 弧的全集记为 E_s , 集合中的元素代表了不同篇章中的同层语言单元之间的语义相似关系。

在 MDF 中, P 弧和 D 弧确定了文本单元在篇章中的地位, S 弧确定了文本单元同其它篇章内部文本单元之间的相似程度, 节点在 x 轴的位置确定了文档集描述的事件流的发生及演变时间。上述几个元素之间互相补充, 互为支撑, 因此, MDF 能够真实地反映多篇章系统的结构。

3 MDF 的建立

3.1 篇章修辞结构的确定

文摘技术中对文本进行修辞结构分析的目的是通过分析各个文本单元之间的相互关系来确定各单元的重要程度。修辞结构理论最早由 Mann^[9] 提出, 并同时给出了由其定义的英文修辞关系系统。对于中文文本, 笔者采用陈清才^[10] 提出的两类修辞关系: (1) 由连接词确定的语句之间的连贯关系, 包括解说、并列、递进、对立、选择、充分、必要、让步、无条件、因果、转折和继承关系; (2) 由文章内容结构线索提供的篇章层次关系, 包括总分、分总和并列关系。本文的篇章修辞结构的建立采用刘挺^[11] 等人的修辞结构树构造方法。为了有效地识别上述两类关系, 我们人工总结了两个线索词词典: 层次结构词典和连接结构词典。分别收录了可以反映文本修辞关系的特征词(词组), 并以此作为识别上述关系的依据。例如: 如果句子中出现线索词“所以”, 则可以认为该词的上下文构成了因果关系。而如果一个段落是由线索词“综上所述”开头, 则该段落之前的若干个段落与之构成了分总关系。

构造篇章的修辞结构必须确定结构中每种关系的权值, 即这种关系在确定各文本单元重要程度中起到的作用。本文采用文献[10]中提到的基于遗传算法的参数优化方法来自动确定各个关系的权值。

如果两个节点 v_i, v_j 之间同时具有连接关系和层次关系, 则采用线性插值方式获得节点之间的最终权值:

$$w_{ij} = \sum_{k=1}^2 \lambda_k \cdot d_k^{ij} \cdot \mu(c_k) \quad (1)$$

式中, $\mu(c_k)$ 为关系 c_k 所对应的权值, 而 $\lambda_k \in [0, 1]$ 是对不同关系类型进行的加权, d_k^{ij} 用以表示关系 c_k 的方向, 如果 c_k 方向由节点 v_i 指向 v_j , 则 $d_k^{ij} = 1$; 否则 $d_k^{ij} = -1$ 。

3.2 S 弧权值的确定

不同文章的片段之间存在着语义相似关系,这种重复信息是多文档自动文摘的重要来源.在 MDF 中,这种语义相似关系用 S 弧来表示.与篇章修辞结构分析不同,由于片段来源于不同文章,因此它们之间的语义相似关系无法利用线索词信息.另一方面,文献[12]提出这种片段间语义相似度计算不能简单地沿用全文相似度计算方法.因此本文采用一种基于多特征融合文本片段相似度计算方法^[13].每一对文本单元之间的语义相似度可由下述逻辑回归公式计算:

$$Y = e^{(\alpha + \beta_1 x_1 + \dots + \beta_8 x_8)} / (1 + e^{(\alpha + \beta_1 x_1 + \dots + \beta_8 x_8)}) \quad (2)$$

式中, α, β 是逻辑回归系数. Y 是文本单元之间的相似度. (x_1, x_2, \dots, x_8) 是从文本单元中抽取的语言特征,笔者采用 1 个词频特征、3 个词性特征、4 个语义特征共 8 个特征来拟合文本单元之间的相似度.

3.3 节点位置的确定

对于 MDF 中的任意一个节点 v, v_y 和 v_z 分别由节点在文章内部的层数以及节点所在的文章位置决定,是一个常数.因此,最终节点位置的确定取决于节点在时间轴上的位置,即 v_x 的值.

3.3.1 文本中的时间信息

时间信息是文本语境理解的重要因素,并与语境中的动作、状态等其它因素密不可分.为了有效地识别并计算文本中的这些时间信息,笔者将承载时间的短语按照功能的不同分解成若干容易识别并且语义单一的“小”的成分,并在此基础上提出了基于规则的时间信息抽取、理解及时间语义的计算方法^[18].句子中的时间信息可用一个可进行语义计算的时间表达式来描述:

$$TE(n) = TO\{TE(n-1)\}, n = \{1, 2, 3\},$$

其中, TE 是递归定义的时间表达式, $TE(0)$ 是原子时间, TO 是时间算子.

3.3.2 节点在 x 轴上的定位规则

我们定义节点在 x 轴的位置为节点所描述的事件(或多个事件)发生的时间(或最晚时间).各级节点在 x 轴的位置确定规则如下.

规则 1. 如果 $u_y = 0$, 则 $u_x = ST(u)$;

规则 2. 如果 $u_y = 3$, 且 $ET(u)$ 已知, 则 $u_x = ET(u)$;

规则 3. 如果 $u_y = 3$, 且 $ET(u)$ 未知, $\exists v \in V$, $ET(v)$ 已知, 并且 $S(u, v) > 0.8$, 则 $u_x = v_x$;

规则 4. 如果 $u_y = 3$, 且 $ET(u)$ 未知, 同时规则 3 的条件不成立, 则 $v'_x < u_x < v_x$, 其中 v, v' 是与 v 同层的上下文节点;

规则 5. 如果 $u_y = 1$ 或者 $u_y = 2$, 且 $\exists v \in V_u$, $\forall u' \in V_u$, 有 $u_x > u'_x$, 则 $V_x = u_x$.

3.3.3 节点的时间关系

节点的时间属性揭示了节点描述的事件的相关程度.如果两个节点同时发生,则这两个节点描述同一事件的概率将大大增加.Allen^[14]在 1984 年提出了两个时间段之间的 13 种基本关系.根据这种分类方法,对于时间轴上的两个节点,我们根据其开始点和终结点的相对位置来确定它们之间的时间关系,并将这种时间关系按照表 1 所示进行量化处理,其中时间系数反映了节点 A, B 描述同一事件的可能性.

表 1 时间关系表示及量化处理表

时间关系	图示	时间系数
A before B		0
B after A		0
A meets B		0
B met-by A		0
A during B		1
B contains A		1
A overlaps B		0
B overlapped A		0
B starts A		1
A started-by B		1
B finishes A		1
A finished-by B		1
A equals B		1

4 基于 MDF 的多文档自动文摘模型

基于 MDF 的多文档自动文摘模型的目标是根据节点之间的关系来选择重要的句子,并按照它们之间的时序关系组成文摘.MDF 中, D 弧决定了节点在同层节点中的重要程度; P 弧决定了节点在上级节点中的地位; S 弧体现了节点与文档集中的其它节点的相关程度.因此,这三方面因素综合考虑才能最终确定节点在文档集中的重要程度.

4.1 节点权值计算

本文通过对每个节点赋予某一权值来标定这个节点的重要程度.从上面的分析可知,MDF 中每个节点具有如下信息:(1) 它在同层节点中的重要程度(D 弧);(2) 它在同文中的上级节点的角色(P 弧);(3) 它与其它节点之间是否相似(S 弧);(4) 它描述的事件发生的时间.节点的权值大小取决于上述因素的融合.根据文献[11]的理论,如果单纯考虑文本修辞关系的影响,一个节点 v 的权值应为

$$a_v = W(u) + P(w, u) + D(w, v) \quad (3)$$

其中, u, w 分别是 v 的上级节点以及同层支配节

点. 另外考虑到当一个节点在 MDF 中的相似节点较多、相似节点在其所在文章中又比较重要的话, 该节点的权值应该相应地增加, 因此节点 v 的最终权值可由下式计算:

$$W(v) = a_v + \frac{1}{n} \sum_{i=1}^n (\theta_{vi} (a_i + S_i)) \quad (4)$$

其中, (S_1, S_2, \dots, S_n) 是 v 与其它 n 个节点的 S 弧权值; (a_1, a_2, \dots, a_n) 是其它所有节点的初始权值. $(\theta_{v1}, \theta_{v2}, \dots, \theta_{vn})$ 分别是 v 与其它 n 个节点之间的时间系数. 综上所述, 融合多个连接信息的节点加权算法描述如下.

算法 1. 自顶向下的节点加权算法.

输入: MDF

输出: 句子节点权值列表

1. 令所有节点的权值为 0; 变量 $i=1$;
2. while $i \leq 4$ do
3. 变量 $s =$ 第 i 层节点数量, $j=1$;
4. while $j \leq s$ do
5. 根据式(3)计算第 j 个节点 v 的初始权值;
6. 根据式(2)计算 v 的最终权值;
7. $j=j+1$;
8. $i=i+1$;
9. 输出句子节点权值列表, 算法结束.

4.2 基于主题聚类的句子节点约简

句子节点权值列表给出了所有句子的权值. 然而不幸的是通过对文摘结果跟踪我们发现, 如果直接采用权值排名前几位的句子作为文摘会产生严重的冗余现象, 原因在于权值相近的句子往往属于同一主题, 而通常情况下, 同一主题的句子只出一个作为代表进入文摘就已经足够了. 因此我们通过聚类的方法进一步对句子节点进行约简.

算法 2. 基于层次聚合聚类聚类的句子节点约简算法.

输入: 待聚类的句子节点集合 H , 阈值 Φ

输出: 类别集合 T

1. 每个句子作为一个类别, 得到初始类别集合 T ;
2. 如果 $\forall t_1, t_2 \in T$, 他们之间的距离均满足下式条件:
 $Dis(t_1, t_2) < \Phi$, 则转步 5;
3. 计算任意两个类别之间的距离:
 $Dis(t_1, t_2) = \max\{Sim(u_i, u_j) \mid u_i \in t_1, u_j \in t_2\}$;
4. 如果存在大于给定阈值 Φ 的类别距离, 合并距离最近的类别, 升级 T , 转步 2;
5. 输出.

算法中类间距离采用最小距离法来计算, 目的在于压缩聚类空间, 生成尽量少的类别, 最大限度地避免冗余问题的出现. 节点 u, v 的距离计算公式如下所示:

$$Sim(u, v) = (1 + \theta_{u,v}) \cdot (S(u, v) + D(u, v)).$$

当 $u \neq v$ 时, $D(u, v)$ 项为 0; $\theta_{u,v}$ 是节点 u, v 之间的时间系数.

5 系统评测

5.1 评测标准

一个理想的文摘系统必须: (1) 包含用户所需的关键信息; (2) 不含冗余信息; (3) 文摘单元之间排列通顺. 因此, 用于分析文摘系统结果的评测标准应该具有评估上述特征的能力. Jones 和 Galliers^[15]定义了两类文摘评估体系: (1) 内部评测. 根据机器文摘同标准文摘的一致程度来评定系统的质量; (2) 外部评测. 评估一个系统在给定任务的执行效果. 其中内部评测的一个问题在于源文档集合中存在与标准文摘句内容相同或相似的句子, 对于一个多文档自动文摘系统来说, 抽取上述任何一个句子都是合理的. Goldstein^[16]采用计算机器文摘句和标准文摘句子之间相似度的方法来测定系统质量, 即机器文摘中存在与标准文摘相似的句子即可, 从某种程度上解决了上述问题. 然而这种方法受到语句相似度计算精度的影响, 很难绝对客观地反映机器文摘的质量.

在本文的评测体系中, 我们采用对评测语料的模糊标注的方法来解决上述问题. 对于标准文摘中的每一个文摘句, 我们作如下定义.

定义 7. 源文档集合中可替换标准文摘句且不能与标准文摘句在文摘中同现的句子称为候选文摘句. 每个候选文摘句根据可替换程度赋予一个取值在 $(0, 1]$ 之间的权值.

定义 8. 标准文摘句同与其相关的所有候选文摘句组成同类句簇, 同属于一个句簇的两个句子称为同类文摘句.

这样, 文摘系统质量评估可用准确率、冗余度和总体质量三项指标来评估, 计算公式如下:

$$precision = (\sum_{i=1}^{k_1} \omega_i) / K,$$
$$redundancy = (\sum_{i=1}^{k_1} (\sum_{j=i+1}^{k_1} \phi(s_i, s_j))) / K,$$
$$total(summary) = precision - redundancy,$$

其中, K 是待评测文摘的句子总数. k_1 是标准文摘的句子在待评测文摘中出现的句子总数, $(\omega_1, \omega_2, \dots, \omega_{k_1})$ 是每个句子的权值, 该权值由上述手工标注方法得到; $\phi(s_i, s_j)$ 是一个二元判别函数, 当 s_i, s_j 为同类文摘句时, $\phi(s_i, s_j) = 1$; 否则为 0.

5.2 文摘结果评测

5.2.1 baseline 系统

baseline 系统包括两个系统：(1)采用人工抽取文摘方法的 upper bound 系统，即人工建立文档集合的 MDF 并根据第 4 章描述的文摘句抽取算法计算每个句子的权值，形成文摘。(2)lower bound 系

统，采用文献[17]提出的多文档自动文摘系统。原因在于该系统同样采用文本单元之间的修辞关系、语义相似关系来最终确定文本单元的权值。与本文不同的是，该方法把上述文本单元加权过程作为多个独立模块串行连接(如图 2 所示)，并没有通过一个整体结构将这些模块有机融合起来。

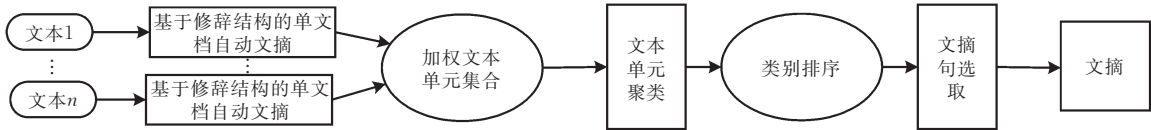


图 2 lower bound 系统流程图

5.2.2 结果及分析

评测语料采用网络收集的 32 组不同主题的真实文档，涉及到体育、金融、突发事件等多个领域。本文的基于 MDF 的方法、upper bound 系统、lower bound 系统的输出结果分别同标准文摘进行比较，结果如表 2~表 4 所示。

表 2			
(a) MDF 方法输出结果			
文摘长度/句	准确率/%	冗余度	总体质量/%
5	70.31	0	70.31
10	68.125	9.37	58.755
20	72.66	7.81	64.85

(b) MDF(不进行时间关系分析)输出结果			
文摘长度/句	准确率/%	冗余度	总体质量/%
5	66.25	1.25	65
10	67.5	13.44	54.06
20	70.25	12.65	57.6

表 3 lower bound 系统输出结果			
文摘长度/句	准确率/%	冗余度	总体质量/%
5	57.7	5.625	52.075
10	55.31	8.44	46.87
20	60.16	9.06	51.1

表 4			
(a) upper bound 系统输出结果			
文摘长度/句	准确率/%	冗余度	总体质量/%
5	88.125	0	88.125
10	90.625	4.68	85.945
20	86.17	5.94	80.23

(b) MDF 中的 P、D 弧采用机器自动识别的结果			
文摘长度/句	准确率/%	冗余度	总体质量/%
5	87.69	0	87.69
10	86.94	4.68	82.26
20	80.34	5.94	74.4

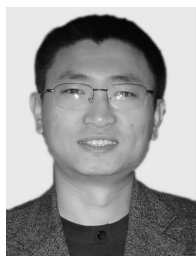
(c) MDF 中的 S 弧采用机器自动识别的结果			
文摘长度/句	准确率/%	冗余度	总体质量/%
5	74.94	0	74.94
10	78.625	8.75	69.875
20	74.68	7.81	66.87

6 结束语

本文提出了一个多文本表示框架 MDF，该框架将文本单元之间的修辞结构信息、语义相似信息以及时序信息通过对文本集合的多层次表示而有机集成在一起。同时，本文在 MDF 基础上提出了文摘句加权、抽取约简以及文摘质量评测等一整套算法。实验结果表明，这种基于 MDF 的多文档自动文摘策略可以获得较高的准确率和较低的冗余度。另外，对比实验同样显示了构建 MDF 的过程中各个关系识别模块精度的不足，这将是笔者今后研究的重点。由于篇幅所限，对于一些 MDF 内部关系分析的具体方法，如时间信息获取及语义计算、文本单元相似度计算等，笔者将另文介绍。

参 考 文 献

- [1] Radev D R et al. Experiments in single and multiple documents summarization using MEAD//Proceedings of the Document Understanding Conference. New Orleans, 2001
- [2] McKeown K, Radev D R. Generating summaries of multiple news articles//Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, 1995: 74-82
- [3] Hardy H et al. Cross-document summarization by concept classification//Proceedings of the Workshop on Text Summarization(DUC 2001). New Orleans, 2001: 65-69
- [4] Boros E et al. A clustering based approach to creating multi-document summaries//Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, LA, 2001: 34-42
- [5] Yi G, Stylios George. A new multi-document summarization system//Proceedings of the Document Understanding Conference. Edmonton, Canada, 2003: 102-109
- [6] Radev D R. A common theory of information fusion from multiple text sources step one: Cross-document structure//Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue. Hong Kong, China, 2000: 74-83
- [7] Zhang Zhu et al. Towards CST-enhanced summarization//Proceedings of the AAAI-2002. Edmonton, Canada, 2002: 439-446
- [8] Zhang Zhu et al. Learning cross-document structural relationships using boosting//Proceedings of the 12th International Conference on Information and Knowledge Management CIKM. New Orleans, Louisiana, USA, 2003: 124-130
- [9] Mann W, Thompson S. Rhetorical structure theory: A theory of text organization. ISI, Los Angeles: Technical Reports ISI/RS-87-190, 1-81, 1987
- [10] Chen Qing-Cai. Research on rough sets based Chinese language modeling and its applications[Ph. D. dissertation]. Harbin Institute of Technology, Harbin, 2003(in Chinese)
- (陈清才. 基于粗集的汉语建模及其应用研究[博士学位论文]. 哈尔滨工业大学, 哈尔滨, 2003)
- [11] Liu Ting, Wang Kai-Zhu. Research on automatic abstracting based on text multi-level dependency structure. Journal of Computer Research & Development, 1999, 36(4): 479-488 (in Chinese)
- (刘挺, 王开铸. 基于篇章多级依存结构的自动文摘研究. 计算机研究与发展, 1999, 36(4): 479-488)
- [12] Hatzivassiloglou V et al. Detecting text similarity over short passages: exploring linguistic feature combinations via machine learning//Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. College Park, Maryland, 1999: 203-212
- [13] Xu Yong-Dong et al. Using multiple features and statistical model to calculate text units similarity//Proceedings of the 4th IEEE International Conference on Machine Learning and Cybernetics. Guangzhou, China, 2005: 3834-3839
- [14] Allen J F. Towards a general theory of action and time. Artificial Intelligence, 1984, 23(2): 123-154
- [15] Jones K S, Galliers J R. Evaluating Natural Language Processing Systems: An Analysis and Review. Berlin: Springer, 1996
- [16] Goldstein J et al. Creating and evaluating multi-document sentence extract summaries//Proceedings of the 9th International Conference on Information and Knowledge Management. Virginia, USA, 2000: 165-172
- [17] Marcu D, Gerber L. An inquiry into the nature of multi-document abstracts, extracts, and their evaluation//Proceedings of the NAACL-2001 Workshop on Automatic Summarization. Pittsburgh, PA, 2001: 1-8
- [18] Xu Yong-Dong, Xu Zhi-Ming, Wang Xiao-Long. Extraction and semantic computing of Chinese textual time information. Journal of Harbin Institute of Technology, 2007, 39(3): 438-442(in Chinese)
- (徐永东, 徐志明, 王晓龙. 中文文本时间信息抽取及语文计算. 哈尔滨工业大学学报, 2007, 39(3): 438-442)



XU Yong-Dong, born in 1974, Ph.D. candidate. His research interests include computational linguistics, automatic summarization.

XU Zhi-Ming, born in 1967, associate professor. His research interests include text retrieval, text mining.

WANG Xiao-Long, born in 1955, professor, Ph. D. supervisor. His research interests include artificial intelligence, natural language processing.

Background

This work is sponsored by key project of National Nature Science Foundation of China (60435020). The project name is "Research of Question-Answering Information Retrieval Technique". In order to improve quality of retrieval system, this project uses multi-document automatic summarization to extract important content from retrieval results and return final answer to users. As the post-processing part

of project, the work in this paper exceedingly affects the result of retrieval system. Before this project, the research term has accomplished relative project of National Natural Science Foundation of China (60373100) named by "Multi-document Automatic Summarization Based on Logistic Frames" which studies Chinese multiple documents automatic summarization for extensive Web information process task.