面向 TDT 的主题相似性计算模型[1]

朱靖波 陈文亮 姚天顺

自然语言处理实验室

东北大学信息学院计算机软件与理论研究所 辽宁 沈阳 110006

Website: Http://www.nlplab.com E-mail: zhujingbo@mail.neu.edu.cn

摘 要：TDT 的研究内容可以分为五个技术任务，本文主要研究第五个技术任务 Linking，即面向 TDT 的事件主题相似性分析技术。研究目的在于力求寻求一种有效的分析技术，针对不同两个文档，识别文档内容所涉及到的事件主题是否一致。分析过程分为两步：(1) 采用 FIFA 模型进行内容主题识别；(2) 采用 LDM 模型进行事件主题相似性计算分析。最后根据实验结果评估两个模型的性能。

关键词：TDT，FIFA 模型，LDM 模型，内容主题识别，事件主题分析

TDT-oriented Topic Similarity Computation Model

Zhu Jingbo Chen Wenliang Yao Tianshun

Natural Language Processing Laboratory

Institute of Compute Software & Theory of Northeastern University, Shenyang, Liaoning, 110006

Abstract: We factor TDT into five technical tasks. This paper focuses on fifth technique-linking detection. An event topic similarity computation technique is studied. Our research works aim to seek an effective analysis technique used to decide whether two documents (or stories) are on the same event topic. There are two steps in the analysis procedure. (1) using FIFA model to identification content topic; (2)using LDM model to compute event topic similarity. At last performances of the two models are evaluated by experiment results.

Keywords: TDT, FIFA Modal, LDM Modal, Content Topic Identification, Event Topic Analysis

1 前言

TDT（Topic Detection and Tracking）最初想法起于 1996 年，从 1999 年开始，连续召开了四届学术会议：TDT1999、TDT2000、TDT2001 和 TDT2002。TDT 的研究工作不同于传统的信息检索、信息抽取、文档分类、信息管理和数据挖掘等技术，主要原因在于 TDT 技术比较关注识别新的事件主题，和获取特定事件主题相关的数据，这一点不同于传统主题类别相关的任务。也就是说，TDT 研究工作中对主题的定义描述不同于传统的主题类别定义，TDT 的主题描述倾向于事件、活动、故事情节等描述，而传统的主题描述倾向于内容主题类别定义，类似于信息的分类，如体育主题、军事主题等。为了方便，本文作如下约定：传统的主题描述简称为内容主题（Content Topic），TDT 中的主题描述简称为事件主题（Event Topic）。

[1] 获得国家自然科学基金和微软亚洲研究院联合资助（No. 60203019）

TDT 的研究人员力求设计一种功能强大、通用、自动学习算法，能够识别和获取人类语言数据的主题结构。这些算法独立于数据的来源、媒介、语种、领域和具体应用。总体来说，TDT 的研究内容可以分为五个技术任务[1]：1）将数据流分割成为多个故事（Segment）；2）寻找属于特定事件主题的所有故事（Tracking）；3）发现新事件主题的所有故事并进行线索化（Detection）；4）发现新事件主题的第一个故事（First Story Detection）；5）确定两个故事涉及的内容是否属于同一个事件主题（Linking）。

本文主要研究第五个技术 Linking，即研究面向 TDT 的事件主题相似性计算技术。第五个技术 Linking 相当于为其它四个技术任务提供了一个基础关键技术。研究目的寻求一种有效的分析技术，针对不同两个文档，识别文档内容所涉及到的事件主题内容是否一致。这一点不同于传统主题识别任务，因为内容主题类别相同的文档（如属于军事主题），所涉及到的事件主题不一定一致（如一篇讨论美军攻打阿富汗、另外一篇讨论美军攻打伊拉克）。因此本文提出的分析技术主要分为两步：

- 首先判断两篇文档的内容主题类别是否一致（如是否都属于军事主题类别？）；
- 然后判断两者所涉及到的事件主题是否一致（如是否都讨论美军攻打伊拉克？）。

针对第一步的研究内容我们可以采用传统的主题分析技术进行分析，本文采用了 FIFA（Feature Identification and Feature Aggregation based）算法进行内容主题类别识别。如果两篇文档的内容主题类别不一致，则判定两者涉及到的事件主题不相似；否则对属于同一内容主题类别的两篇文档进行事件主题分析，根据分析结果判定两者涉及到的事件主题是否相似。

2 内容主题识别模型 FIFA

目前国内外很多学者对内容主题识别技术进行深入研究，提出了很多方法，取得了一些研究成果。这些方法可以分为三种：基于统计的方法[2][3][4]、基于知识的方法[5]和结合两者混合的方法[6][7]。本文认为，主题分析和文本层次结构分析除了充分利用传统的语法语义知识库外，还需要充分利用世界背景知识或领域知识。

下面简单介绍一下朱靖波和姚天顺[8]提出的一种文本内容主题识别技术：FIFA 算法，其中主要研究了充分利用大规模的领域知识库，将基于词汇层的分析技术提升到领域知识的计算层面，实现文本内容主题识别。

FIFA 算法描述如下：

Step1 对于输入的文本进行分词和词性标注；

Step2 利用主题特征识别技术计算文本主题特征分布，生成文本的主题特征集合；

输入：分词和词性标注后的文本 T。

1） 基于特征词典的特征项识别和标注过程：该方法主要思想是利用特征词典中特征项来完成文本中包含的特征项识别和标注。一旦发现文本中包含特征词典中描述的特征项，就从特征词典中检索出该特征项的属性作为标注，同时将该特征项表达的主题特征标注为该特征项的领域属性。

2） 基于规则的特征项识别和标注过程：主要包括两步：① 采用统计的方法，从文本中自动提取出高频字串作为分析对象，我们的系统默认为长度大于二个词汇，出现频度大于二次。② 采用基于规则的方法，对自动提取的高频字串进行结构分析，根据字串的结构和中心名词的属性，猜测该高

频字串的领域属性，将领域属性标注为该高频字串表达的主题特征。

3）根据特征项的词典属性、频率和在文本中的位置，计算文本 T 的主题特征分布；

输入：文本的主题特征集合 ψ。

Step3 利用集聚公式计算文本主题分布，并根据主题权值大小进行排序；

输入：被分析文本的主题特征集合 ψ。

1） 从集聚公式库中获取一条集聚公式 ξ_i，其中 ξ_i 是主题 t_i 的集聚公式；

2） 基于文本主题特征集合 ψ 中的参数，根据积聚公式 ξ_i 计算主题 t_i 的权值；

3） 如果集聚公式库还存在其它主题的公式，则 Step3：1），否则执行下一步；

4） 默认文本主题特征集合 ψ（$=\{(f_i, p(f_i))\}$）中的 f_i 为一个主题，$p(f_i)$ 为主题 f_i 的权值。根据主题权值的大小将主题 t_i 和主题 f_i 进行排序；

输出：选择最大权值的主题作为文本的内容主题标注。

算法 1　FIFA 算法描述

特征词汇领域属性词典（下文简称特征词典）是一个领域知识库。自从 1996 年至今，我们开发的特征词典中包含了 30 多万条词条，全部根据真实语料，首先进行人工分类，然后采用机器辅助人工校对的方法构造的。我们认为充分利用传统语法语义知识和领域知识是一种提高内容主题分析的有效途径。特征词典的数据结构包括：词形、词类、语义分类、语义特征、地域属性和主题特征（本文也称作领域属性）等信息。如：两岸会谈，N(),,会谈,事情(两岸会谈),,省市(台湾)国名(中国),主题特征(台湾问题)主题特征(政治)。

根据特征项 ft_i 的频度和位置信息计算该特征项的权值，计算公式如下：

$$p(ft_i) = \frac{freq(ft_i) + N_{title} + 0.5 \times N_{begin} + 0.5 \times N_{end}}{\sum freq(ft_i)} \qquad (1)$$

其中，$p(ft_i)$：表示特征项 ft_i 的权值；$freq(ft_i)$：表示特征项 ft_i 的的频度；N_{title}：表示特征项 ft_i 在标题中出现的次数；N_{begin}：表示特征项 ft_i 在段首句中出现的次数；N_{end}：表示特征项 ft_i 在段尾句中出现的次数；$\sum freq(ft_i)$：表示文本中所有特征项出现的次数。

说明：研究中发现，处在文本中不同位置的特征项表达主题特征的能力有所不同，因此在特征项的权值计算中，我们考虑了这个因素，通过给定经验值的方法来解决。正如公式（1）中的 N_{title} 的系数为 1.0，N_{begin} 的系数为 0.5，N_{end} 的系数为 0.5。

从每个特征项的词典属性中可以获得该特征项的领域属性，该领域属性就是该特征项所表达的主题特征。通过所有表达同一个主题特征的特征项的权值进行总和，就可以计算出该主题特征的权值。主题特征的权值越高，表明表达该主题特征的特征项越多，反映该文本的主题倾向性越强。主题特征 f_i 的权值 $p(f_i)$ 的计算公式如下：

$$p(f_i) = \sum_{w_j \in f_i} p(w_j) \qquad (2)$$

其中：$p(f_i)$ 表示主题特征 f_i 的权值；$p(t_j)$ 表示特征项 w_j 的权值；$w_j \in f_i$ 表示表达主题特征 f_i 的所有特征项 w_j。根据公式（2）的计算，就可以构造：

文本的主题特征分布向量 ψ_t：$\psi_t = \{<f_{t1}, p(f_{t1})>, <f_{t2}, p(f_{t2})>, \ldots, <f_{ts}, p(f_{ts})>\}$ 和

段落的主题特征分布向量 ψ_p：$\psi_p = \{<f_{p1}, p(f_{p1})>, <f_{p2}, p(f_{p2})>, \ldots, <f_{pm}, p(f_{pm})>\}$。

主题 t_i 主题特征的集聚公式 ξ_i 定义如下：

$$\xi_i : p(t_i) = \sum_{j=1}^{n} p(f_j) \times \mu(f_j) \qquad (3)$$

其中：$p(t_i)$表示主题 t_i 的权值；f_j表示属于主题 t_i 的主题特征；$p(f_j)$表示主题特征 f_j 的权值；$\mu(f_j)$表示在集聚公式中主题特征 f_j 的系数。

其中主题特征集聚公式中的主题特征和系数可以从事先分好类的训练语料中通过自动构造的方法自动获取。在我们开发的系统中自动构造了 105 个主题的集聚公式，其中包括体育、旅游、金融、性保健、黄色、法轮功等等，将这 105 个主题特征集聚公式按照固定的格式存放到一个文件中，本文称之为主题特征集聚公式库。

3. 事件主题相似性计算模型 LDM

3.1 事件主题的描述

事件主题（Event Topic）的描述不同于传统的内容主题（Content Topic）类别的描述，涉及到事件发生的地点（Where）、事件发生的时间（When）、事件发生的对象（Object）等。本文给出一种事件主题的描述框架式结构，其中包括的事件主题的属性描述如下：

EVENT_NO：事件主题的编号，由系统自行设置；EVENT_NAME：事件主题的名称；EVENT_WHEN：事件发生的事件范围；EVENT_WHERE：事件发生的地点；EVENT_OBJECT：事件涉及到的主体对象，包括人、机构、组织等；EVENT_CAUSE：事件发生的原由，可以为空；EVENT_RESULT：事件发生产生的后果，可以为空；EVENT_DESCRIPTION：事件的辅助描述，可以为空。

3.2 LDM 模型

目前很多研究人员提出了一些分析技术，基本思想在于构造文档（或故事）的特征向量，采用 Cosine 相似计算方法进行计算两篇文档（或故事）的事件主题相似性。这种采用传统内容主题分析技术进行事件主题相似性计算，主要依赖于基于词汇特征项的统计分析技术，本文认为具有很大的局限性。

本文提出了一种事件主题相似性计算模型 LDM（Linking Detection Modal），基本思想在于构造两篇文档（或故事）的主题特征概率分布矩阵，通过计算两个主题特征矩阵的相似性来判断文档内容所涉及到的事件主题相似性。本文的方法是基于领域知识和主题特征的计算层面。

LDM 模型的具体实现算法如下：

1）分别构造两篇文档的主题特征概率分布矩阵 T_1 和 T_2：

$$T_1 = \begin{pmatrix} \psi_{t11} \\ \psi_{t12} \\ \cdots \\ \psi_{t1n} \end{pmatrix} = \begin{pmatrix} p(f_{111}) & p(f_{112}) & \cdots & p(f_{11m}) \\ p(f_{121}) & p(f_{122}) & \cdots & p(f_{12m}) \\ & & & \\ p(f_{111}) & p(f_{112}) & \cdots & p(f_{11m}) \end{pmatrix}$$

$$T_2 = \begin{pmatrix} \Psi_{p21} \\ \Psi_{p22} \\ \cdots \\ \Psi_{p2n} \end{pmatrix} = \begin{pmatrix} p(f_{211}) & p(f_{212}) & \cdots & p(f_{21m}) \\ p(f_{221}) & p(f_{222}) & \cdots & p(f_{22m}) \\ & & & \\ p(f_{211}) & p(f_{212}) & \cdots & p(f_{21m}) \end{pmatrix}$$

2）采用 Kullbacj-Leibler 距离函数 ζ 进行相似计算文档主题特征概率分布矩阵 T_1 和 T_2 的相似性，计算公式为：$Sim(T_1,T_2)=\sqrt{\exp(-\varsigma(T_1,T_2))}$。距离函数 ζ 见下文公式（4）。

3）设置相似阀值 θ，如果 $Sim(T_1,T_2)<\theta$，则判定事件主题不相似。

4）如果相似度大于阀值 θ 的话，进行基于事件主题描述框架的匹配机制。如果两篇文档都包含描述框架所提供的关键词，则判定该两篇文档的事件主题相似，否则不相似。

3.3 基于 Kullback-Leibler 距离的距离函数

首先定义 o_1 和 o_2 为两个不同的对象，CT 为局部上下文，c∈CT 表示某一具体局部上下文。P_{o1} 和 P_{o2} 分别表示对象 o_1 和 o_2 的概率分布。Kullback-Leibler[9]距离公式定义如下：

$$D(P_{o1} \| P_{o2}) = \sum_{c \in CT} P(c \mid o_1) \log \frac{P(c \mid o_1)}{P(c \mid o_2)}$$

本文利用 $D(P_{o1}\|P_{o2})+ D(P_{o2}\|P_{o1})$来计算两个不同概率分布 P_{o1} 和 P_{o2} 的距离，距离函数 ζ (P_{o1},P_{o2})如下：

$$\begin{aligned} &\varsigma(P_{o1},P_{o2}) = D(P_{o1} \| P_{o2}) + D(P_{o2} \| P_{o1}) \\ &= \sum_{c \in CT} P(c \mid o_1) \log \frac{P(c \mid o_1)}{P(c \mid o_2)} + \sum_{c \in CT} P(c \mid o_2) \log \frac{P(c \mid o_2)}{P(c \mid o_1)} \\ &= \sum_{c \in CT} (P(c \mid o_1) \log \frac{P(c \mid o_1)}{P(c \mid o_2)} + p(c \mid o_2) \log \frac{P(c \mid o_2)}{P(c \mid o_1)}) \\ &= \sum_{c \in CT} ((P(c \mid o_1) - p(c \mid o_2)) \log \frac{P(c \mid o_1)}{P(c \mid o_2)}) \\ &= \sum_{c \in CT} ((P(c \mid o_1) - p(c \mid o_2))(\log P(c \mid o_1) - \log P(c \mid o_2)) \end{aligned} \quad (4)$$

4 FIFA+LDM 模型实验

为了验证基于 FIFA+LDM 模型的事件主题相似性分析技术的有效性，本文构造了两个测试集 A1 和 A2，分别包含 200 篇新闻文档，包括三个内容主题类别：体育、军事和娱乐。选取了六个事件：伊拉克战争、阿富汗战争、LG 围棋比赛、乒乓球比赛、英雄电影和艺人裸照事件，关于每个事件的新闻文档选取了 40 篇，每个测试集分别包含 20 篇，则共有 20*20*6=2400 个测试对是符合条件的。分别从两个测试集中获取一个新闻文档，可以构造一个测试对，这样可以构造 200*200=40000 个测试对，每个测试对包括两篇新闻文档。实验结果显示，相似阀值与正确率、召回率和 F 评价的关系如下图所示。

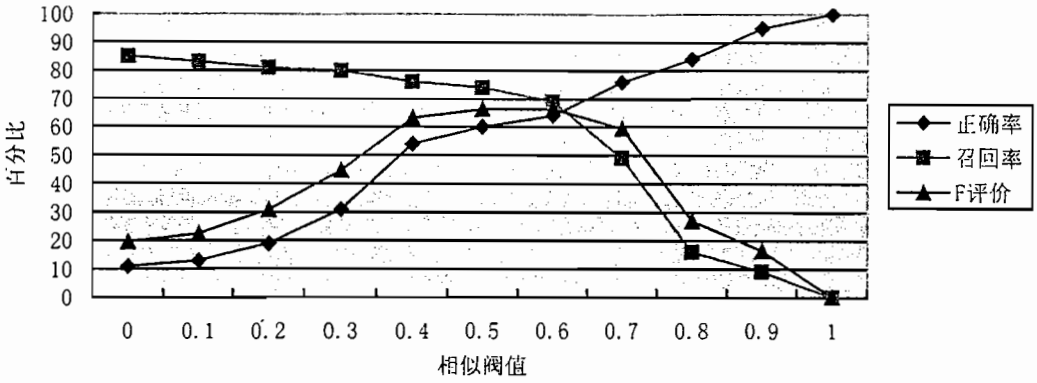

图 1 相似阀值与正确率和召回率的关系

从图 1 中可以发现，当相似阀值 θ 设置为 0.6 时，系统的总体性能达到最佳。其中 F 评价值达到最大（0.664），正确率为 64%，召回率为 69%，丢失率为 31%。

5 结束语

本文提出了一种事件主题相似性计算技术，首先采用 FIFA 模型进行内容主题识别，在内容主题类别相同的情况下，继续采用 LDM 模型进行事件主题相似性计算。该方法的创新之处在于充分利用了领域知识库，将传统分析技术从词汇层提升到领域知识和主题特征计算层面，采用两步分析机制，有效提高了事件主题分析的性能。下一步我们将采用该分析技术应用于 TDT 的其他四个技术任务研究工作中。

参考文献：

[1] Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation. Wayne, C., Language Resources and Evaluation Conference (LREC) 2000, pages 1487-1494.

[2] H.P.Luhn. A statistical approach to mechanized encoding and searching of literary information. IBM Journal, p309-17, October 1957

[3] H.P.Edmundson. New methods in automatic extracting. Journal of the ACM. 16(2):264-85.1969

[4] Gerard Salton, James Allan, Chris Buckley, and Amit Singhal. Automatic analysis, theme generation, and summarization of machine-readable texts. Science. 264:1421-26. June 1994

[5] Wendy Lehnert and C. Loiselle. An introduction to plot unit. In David Waltz, Semantic Structures-Advances in Natural Language Processing. Lawrence Erlbaum Associates. Hillsdale. New Jersey, p88-111, 1989

[6] Marti A. Hearst. Context and Structure in Automated Full-Text Information Access. PhD thesis. Computer Science Division. University of California at Berkeley. California. April 1994

[7] Eduard Hovy, and Chin-Yew Lin. Automated text summarization in SUMMARIST. In ACL/EACL97 Workshop on Intelligent Scalable Text Summarization. p18-24.1997

[8] 朱靖波，姚天顺，基于 FIFA 算法的文本分类，中文信息学报，Vol16，No3，2002

[9] Kullback. Solomon: Information Theory and Statistics. John Wiley and Sons, New York. 1959