ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.9, September 2008, pp.2276−2284
DOI: 10.3724/SP.J.1001.2008.02276

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于信息论的潜在概念获取与文本聚类*

李晓光[1+], 于 戈[2], 王大玲[2], 鲍玉斌[2]

[1](辽宁大学 信息学院,辽宁 沈阳 110036)

[2](东北大学 信息科学与工程学院,辽宁 沈阳 110004)

Latent Concept Extraction and Text Clustering Based on Information Theory*

LI Xiao-Guang[1+], YU Ge[2], WANG Da-Ling[2], BAO Yu-Bin[2]

[1](Department of Information, Liaoning University, Shenyang 110036, China)

[2](Department of Information Science and Engineering, Northeastern University, Shenyang 110004, China)

+ Corresponding author: E-mail: xgli@lnu.edu.cn

Li XG, Yu G, Wang DL, Bao YB. Latent concept extraction and text clustering based on information theory. *Journal of Software*, 2008,19(9):2276−2284. http://www.jos.org.cn/1000-9825/19/2276.htm

Abstract: To emphasize the fuzzy relation among words, latent concepts, text and topics, an information theory based approach to latent concept extraction and text clustering is proposed. Latent concept variable and topic variable are introduced to reveal such relation, and a global objective function is defined in the theme of rate-distortion theory. An anneal-like algorithm is designed to extract the hierarchical tree of latent concept, and to group the texts under corresponding concept hierarchy at the same time. Furthermore, it determines the number of concept and text clustering result with a concept selection method based on minimal description length criteria. It is a soft co-clustering method and outperforms the ones based on the word space, and current text hard co-clustering method based on latent concept by experiments.

Key words: latent concept; topic; text clustering; Information theory

摘 要: 针对词、潜在概念、文本和主题之间的模糊关系,提出一种基于信息论的潜在概念获取与文本聚类方法.方法引入了潜在概念变量和主题变量,根据信息论中熵压缩编码理论,定义了一个全局目标函数,给出一种类似于确定性退火算法的求解算法,用以获得概念层次树以及在不同层次概念上的文本聚类结果,是一种双向软聚类方法.方法通过基于最短描述长度原则的概念选择方法,最终确定概念个数和对应的文本聚类结果.实验结果表明,所提出的方法优于基于词空间的文本聚类方法以及双向硬聚类方法.

关键词: 潜在概念;主题;文本聚类;信息论

中图法分类号: TP18 **文献标识码**: A

文本聚类方法大多以词作为基本特征,然而由于词空间的高维性、稀疏性和相关性等特点,聚类质量和效率并不能令人满意[1].特征选取尽管可以减少特征空间的维数,降低稀疏性,却无法解决特征之间语义相关性的问题.人们希望能够将具有描述相同语义能力的词作为一个语义单位并以此为文本的基本特征,从而获得低

* Supported by the National Natural Science Foundation of China under Grant Nos.60703068, 60573090 (国家自然科学基金)
Received 2006-12-28; Accepted 2007-08-03

维、独立的特征空间.这里统一称这种语义单位为潜在概念.一般来说,词、潜在概念、文本和主题有以下关系:

(1) 词与潜在概念相关.由于词的多义性,不同概念的交集可以不为空,也就是说,不同词可以表达相同的概念,同一个词又可以表达不同的概念,并且不同词对同一个概念的表达能力也不同.

(2) 文本与主题相关.文本内容描写往往是围绕某个主题来展开的,不同的文本可以具有相同的主题,而相同的文本可以包含多个不同的主题.

(3) 潜在概念与主题相关,也就是说,一个主题可以包含多个概念,并且一个概念可以属于多个主题.

针对这种模糊关系,本文提出了一种基于信息论的潜在概念获取与文本聚类方法,它引入了潜在概念变量和主题变量,并以概率形式表示词、潜在概念、文本和主题之间的模糊关系.根据信息论中熵压缩编码理论,定义了一个全局目标函数,给出了一个类似于确定性退火算法的求解算法,在获得概念层次树的同时,也获得了在不同层次概念上文本基于主题的聚类结果,是一种双向软聚类方法.另外,本文通过一个基于最短描述长度原则的概念选择方法,来确定所获取的概念个数和对应的文本聚类结果.实验结果表明,本文提出的方法要优于基于词空间的文本聚类方法以及双向聚类方法.

本文第 1 节为相关工作.第 2 节提出问题并描述本文提出的模型.第 3 节根据模型定义,给出具体求解方法.第 4 节给出实验比较结果.第 5 节是全文的结论.

1 相关工作

目前,潜在概念获取与文本聚类主要有以下两种方法:(1) 基于数学分析的方法,如潜在语义分析[2]和基于概率的潜在语义分析[3,4],该类方法认为,文档与词通过一层潜在变量相互关联,如文献[4]中的潜在主题变量,潜在变量不同取值对应着不同的词与文本生成模型.与其不同,本文的模型同时涉及主题变量和潜在概念变量两个层次,词与文本通过这两个层次相互关联,但值得注意的是,这里,主题变量和潜在概念变量没有作为产生词与文本的潜在模型,而是作为词与文本编码后对应的码表,通过求解全局目标函数获得潜在概念和文本基于主题的聚类结果.(2) 基于词聚类分析的方法,如概念检索[5]、概念词链[6]、基于类分布的词聚类[7]、双向聚类[8],然而,由于词空间的高维性和稀疏性,聚类结果往往出现偏斜结果,其概念获取质量很低[1].

本文在设计目标函数时,基本思想来源于“信息瓶颈(information bottleneck,简称 IB)”[9].文献[10,11]提出一种基于统计分布的词聚类方法,并在此基础上提出了信息瓶颈概念.文献[12]根据信息瓶颈的思想,提出了多元信息瓶颈(multivariate information bottleneck).与对称 IB 模型不同,本文从概念获取和主题聚类角度出发,模型的目标函数的确定和随机变量之间的依赖关系,以及相应的目标函数求解都与之不同,而后者只是定义了一个理论框架,并没有涉及到具体的应用,在确定目标函数时,实际上是只考虑了最大化概念与主题、概念与文本和主题与词之间的互信息.文献[13]提出一种基于信息论的文本和词的双向硬聚类方法,其目标函数设计思想类似于 IB,但只考虑了信息失真,本文目标函数同时考虑了信息失真和信息压缩两个方面,是一种软聚类方法.

另外,基于聚类分析获取概念的方法中,很重要的一个问题就是如何确定概念的个数,如采用人工估计的方法、交叉验证的方法等等,然而这些方法都存在效率比较低下的问题.本文根据最短描述长度原则设计了一个选取更泛化的概念的方法,尽管该方法只是得到 MDL 的局部最优解,但实验结果表明,更为泛化的概念可以取得很好的聚类结果,并且可以获得更低维的概念空间.

2 模型描述

对文本集合 $\boldsymbol{D}$,$|\boldsymbol{D}|=n$,其词集合 $\boldsymbol{W}$,$\boldsymbol{W}=\{w|w\in d,d\in\boldsymbol{D}\}$且$|\boldsymbol{W}|=m$.设文本集合 $\boldsymbol{D}$ 中含有 k_c 个潜在概念和 k_t 个主题,概念集合表示为 $\boldsymbol{C}$,主题集合表示为 $\boldsymbol{T}$.本文假设文本集合中主题个数 k_t 为已知,而概念个数 k_c 为未知.

为了体现词、潜在概念、文本和主题之间的关系,对词 $w\in\boldsymbol{W}$ 和潜在概念 $c\in\boldsymbol{C}$ 的关系以条件概率 $P(c|w)$ 来表示,即 w 属于 c 的概率.那么,潜在概念 c 可以用向量方式表达,即 $V(c)=(w_1,P(w_1|c);\ldots;w_i,P(w_i|c);\ldots;w_m,P(w_m|c))$,其中 $P(w_i|c)$为给定概念 c 下词 w_i 的条件概率,可解释为给定概念 c 产生词 w_i 的可能性.同样,对文本 $d\in\boldsymbol{D}$ 和主题 $t\in\boldsymbol{T}$ 的关系以条件概率 $P(t|d)$来表示,即文本 d 属于主题 t 的概率,对潜在概念 $c\in\boldsymbol{C}$ 和主题 $t\in\boldsymbol{T}$ 的关系以

条件概率 $P(t|c)$来表示.根据贝叶斯规则,概念获取与文本聚类可转换为求解概率 $P(c|w),\forall c\in \boldsymbol{C},w\in \boldsymbol{W}$ 和 $P(t|d),\forall d\in \boldsymbol{D},t\in \boldsymbol{T}$,以及概念个数 k_c 估计问题.

如果将 $\boldsymbol{W}$ 和 $\boldsymbol{D}$ 作为源字母表,$\boldsymbol{C}$ 和 $\boldsymbol{T}$ 为相对应的码表,则 $\boldsymbol{W}$ 与 $\boldsymbol{C}$ 之间以及 $\boldsymbol{D}$ 与 $\boldsymbol{T}$ 之间的映射关系可视为熵压缩编码,那么我们希望压缩程度在越大越好的同时,潜在概念和主题之间应该尽可能地保持原有词与文本之间的信息.设随机变量 W,C,D 和 T,分别代表词、潜在概念、文本和主题,其取值范围分别为 $\boldsymbol{W},\boldsymbol{C},\boldsymbol{D}$ 和 $\boldsymbol{T}$.定义随机变量之间的依赖关系: $C\leftrightarrow W\leftrightarrow D\leftrightarrow T$,即 C 和 T 分别只与 W 和 D 相关,W 与 D 相关.对文本集合 $\boldsymbol{D}$,联合概率分布 $P(W,D)$通过最大似然估计获得.本文在衡量信息压缩和信息失真的度量时采用了互信息.$I(W;C)$和 $I(D;T)$分别衡量了 $\boldsymbol{W}$ 和 $\boldsymbol{D}$ 经过编码后映射到 $\boldsymbol{C}$ 和 $\boldsymbol{T}$ 后的信息压缩程度,则信息压缩 $CR=I(W;C)+I(D;T)$.$I(W;D)$和 $I(C;T)$ 分别衡量了词与文本之间和概念与主题之间的相互信息,则信息失真 $DT=I(W;D)-I(C;T)$.目标函数由两方面确定:1) 信息压缩 CR 最大化.信息压缩越大,则源字母与码字之间依存性越低,互信息越小;2) 信息失真 DT 最小化.对于源字 W 与 D 的互信息 $I(W;D)$,经过编码压缩后,码字 C 与 T 的互信息与 $I(W;D)$相比,损失越小越好,由于对给定的文本集合 $\boldsymbol{D}$,$I(W;D)$为常值,则求解信息失真最小化等同于最大化 $I(C;T)$.利用拉格朗日乘数法,将求解信息压缩最大化和信息失真最小化转化为求解下式,其中β为拉格朗日乘子.

$$\text{Min}\quad F=I(W;C)+I(D;T)-\beta\cdot I(C;T) \tag{1}$$

并满足约束:

$$\sum_{c\in C}P(c\mid w)=1,\forall w\in W \tag{2}$$

$$\sum_{t\in T}P(t\mid d)=1,\forall d\in D \tag{3}$$

3 模型求解

对于目标函数 F,β可以看作用于平衡信息压缩和信息失真的参数.如果将式(1)看作自由能函数,β为温度的倒数,则 F 的求解过程类似于确定性退火方法.根据 RD 理论,当 F 取极值时,$\beta=dCR/dDT$,则当β由 0 逐渐增加时,目标函数由侧重信息压缩逐渐转向侧重于信息失真,当β增至产生“相变”的“临界温度”时,某些概念产生分裂,此时,确定性退火算法保证了优化搜索尽可能地接近斜率为β的 RD 曲线[14].下面,首先给出给定β下求解目标函数的收敛方法,然后给出潜在概念个数的选择方法作为退火算法的终止条件,最后,为具体算法设计及相关讨论.另外,为了表示简洁,本文中 $P(c,t)\equiv P(C=c,T=t)$,其他概率表达式依此类推.

3.1 给定β下F求解

根据给定的变量依赖条件,有如下性质:

性质 1. 给定随机变量依赖关系 $C\leftrightarrow W\leftrightarrow D\leftrightarrow T$ 以及 $P(W,D)$,则有

$$P(t\mid w)=\sum_{d\in D}P(t\mid d)P(d\mid w);\forall t\in T,w\in W \tag{4}$$

$$P(c\mid d)=\sum_{w\in W}P(c\mid w)P(w\mid d);\forall c\in C,d\in D \tag{5}$$

$$P(c,t)=\sum_{d\in D}P(t\mid d)P(c\mid d)P(d)=\sum_{w\in W}P(c\mid w)P(t\mid w)P(w);\forall c\in C,t\in T \tag{6}$$

定理 1. 对给定的β 以及联合概率分布 $P(W,D)$,当目标函数 F 取极值时,必须满足条件:

$$P(t\mid d)=\frac{P(t)}{Z(d)}\mathrm{e}^{\beta\sum_{c\in C}P(c|d)\log\left(\frac{P(c,t)}{P(c)P(t)}\right)},Z(d)=\sum_{t\in T}P(t)\mathrm{e}^{\beta\sum_{c\in C}P(c|d)\log\left(\frac{P(c,t)}{P(c)P(t)}\right)};\forall t\in T,d\in D \tag{7}$$

$$P(c\mid w)=\frac{P(c)}{Z(w)}\mathrm{e}^{\beta\sum_{t\in T}P(t|w)\log\left(\frac{P(c,t)}{P(c)P(t)}\right)},Z(w)=\sum_{c\in C}P(c)\mathrm{e}^{\beta\sum_{t\in T}P(t|w)\log\left(\frac{P(c,t)}{P(c)P(t)}\right)};\forall c\in C,w\in W \tag{8}$$

证明:求解目标函数 F 是一个有约束极值问题,利用拉格朗日乘数法将式(1)的求解化为求解函数:

$$F' = F + \lambda(w)\left(\sum_{c\in C} P(c\mid w) - 1\right) + \lambda(d)\left(\sum_{t\in T} P(t\mid d) - 1\right),$$

则 F'函数取极值时必须满足:

$$\frac{\partial F'}{\partial P(t\mid d)} = \frac{\partial I(W;C)}{\partial P(t\mid d)} + \frac{\partial I(D;T)}{\partial P(t\mid d)} - \beta\frac{\partial I(C;T)}{\partial P(t\mid d)} + \lambda(d) = 0 \tag{9}$$

$$\frac{\partial F'}{\partial P(c\mid w)} = \frac{\partial I(W;C)}{\partial P(c\mid w)} + \frac{\partial I(D;T)}{\partial P(c\mid w)} - \beta\frac{\partial I(C;T)}{\partial P(c\mid w)} + \lambda(w) = 0 \tag{10}$$

结合性质 1,有

$$\frac{\partial I(W;C)}{\partial P(t\mid d)} = 0 \tag{11}$$

$$\frac{\partial I(D;T)}{\partial P(t\mid d)} = P(d)\log P(t\mid d)/P(t)) \tag{12}$$

$$\frac{\partial I(C;T)}{\partial P(t\mid d)} = P(d)\sum_{c} P(c\mid d)\log P(c,t)/P(c)P(t) \tag{13}$$

将式(11)~式(13)代入式(9)可得:

$$\frac{\partial F'}{\partial P(t\mid d)} = 0 + P(d)\lg P(t\mid d)/P(t) - \beta P(d)\sum_{c} P(c\mid d)\log P(c,t)/P(c)P(t) + \lambda(d) = 0$$

$$\Rightarrow P(t\mid d) = \lambda'(d)P(t)\mathrm{e}^{\beta\sum_{c\in C} P(c|d)\log\left(\frac{P(c,t)}{P(c)P(t)}\right)}, \lambda'(d) = \mathrm{e}^{\lambda(d)}.$$

因为 $P(t\mid d)$ 须满足约束(3),则

$$\lambda'(d) = \left(\sum_{t\in T} P(t)\mathrm{e}^{\beta\sum_{c\in C} P(c|d)\log\left(\frac{P(c,t)}{P(c)P(t)}\right)}\right)^{-1},$$

式(7)可证.式(8)同理可证. □

由于式(7)和式(8)中的 $P(c,t)$,$P(t|w)$和 $P(c|d)$隐含了 $P(t|d)$和 $P(c|w)$,没有解析解.类似于 Blahut-Arimoto 算法,这里给出目标函数 F 数值解求解算法.

定理 2. 在变量 $P(C|w)=(P(c|w)|c\in\boldsymbol{C})$,$\forall w\in\boldsymbol{W}$ 和 $P(T|d)=(P(t|d)|\ t\in\boldsymbol{T})$,$\forall d\in\boldsymbol{D}$ 以及变量 $P(c,t)$,$P(t|w)$和 $P(c|d)$,$\forall c\in\boldsymbol{C}$,$w\in\boldsymbol{W}$,$d\in\boldsymbol{D}$,$t\in\boldsymbol{T}$ 上,目标函数 F 分别下凸,而在所有变量上非凸.

证明:对 $P(C|w)$来说,因为 $I(C;W)$可改写为

$$I(C;W) = \sum_{w\in W} P(w)\underbrace{\sum_{c\in C} P(c\mid w)\log\left(\frac{P(c\mid w)}{P(c)}\right)}_{(a)},$$

其中(a)部分为 $P(C|w)$的下凸函数,那么下凸函数的线性组合也为下凸函数,并且 F 的其他部分与 $P(C|w)$无关,所以 F 对 $P(C|w)$为下凸函数.对 $P(T|d)$,$P(c,t)$,$P(t|w)$来说,同理可证.很明显,F 在所有的变量上为非凸函数. □

定理 3. 对给定的β以及联合概率分布 $P(W,D)$,则迭代过程:

$$P^{i+1}(t\mid d) = \frac{P^i(t)}{Z^{i+1}(d)}\mathrm{e}^{\beta\sum_{c\in C} P^i(c|d)\log\left(\frac{P^i(c,t)}{P^i(c)P^i(t)}\right)} \tag{14}$$

$$P^{i+1}(c\mid w) = \frac{P^i(c)}{Z^{i+1}(w)}\mathrm{e}^{\beta\sum_{t\in T} P^i(t|w)\log\left(\frac{P^i(c,t)}{P^i(c)P^i(t)}\right)} \tag{15}$$

$$P^{i+1}(t\mid w) = \sum_{d\in D} P^{i+1}(t\mid d)P(d\mid w) \tag{16}$$

$$P^{i+1}(c\mid d) = \sum_{w\in W} P^{i+1}(c\mid w)P(w\mid d) \tag{17}$$

$$P^{i+1}(c,t)=\sum_{d\in D}P^{i+1}(t\mid d)P^{i+1}(c\mid d)P(d)=\sum_{w\in W}P^{i+1}(c\mid w)P^{i+1}(t\mid w)P(w) \tag{18}$$

$$P^{i+1}(c)=\sum_{t\in T}P^{i+1}(c,t) \tag{19}$$

$$P^{i+1}(t)=\sum_{c\in C}P^{i+1}(c,t) \tag{20}$$

当 $i\to\infty$时,F 收敛于局部最优.

证明:首先,明显目标函数 F 有下界,其次将 F 分别看作变量 $P(T|d)$和 $P(C|w)$,以及变量 $P(c,t)$,$P(t|w)$和 $P(c|d)$的函数,则当 F 为 $P(T|d)$的函数时,根据定理 1,其取极值并满足标准化约束时,$P(T|d)$的每个分量必须满足式(7).当 F 为 $P(C|w)$的函数时,则 $P(C|w)$的每个分量必须满足式(8).而当 F 为 $P(c,t)$的函数时,可改写为

$$F=I(C;W)+I(T;D)-\beta\sum_{c\in C}\sum_{t\in T}\sum_{d\in D}P(t\mid d)P(c\mid d)P(d)\log\big(P(c,t)/P(c)P(t)\big),$$

利用拉格朗日乘数法可得:

$$P(c,t)=\sum_{d\in D}P(t\mid d)P(c\mid d)P(d).$$

同理可证,

$$P(c,t)=\sum_{w\in W}P(c\mid w)P(t\mid w)P(w),\ P(c\mid d)=\sum_{w\in W}P(c\mid w)P(w\mid d),$$

以及

$$P(t\mid w)=\sum_{d\in D}P(t\mid d)P(d\mid w).$$

从上述证明可得式(14)~式(18)分别为目标函数 F 在变量 $P(T|d)$和 $P(C|w)$以及变量 $P(c,t)$,$P(t|w)$和 $P(c|d)$上取极值的条件,而式(19)和式(20)为 $P(c,t)$的边缘概率.根据定理 2,F 在每个变量上为下凸函数而在所有变量上非凸,那么当更新某个变量(固定其他变量)时均使得目标函数 F 在该变量上的投影取最小值或者保持不变,并且又由于 F 有下界,则当 $i\to\infty$时,F 收敛于局部最优. □

3.2 概念个数k_c确定

本文概念个数 k_c 确定利用了最小描述长度(minimum description length,简称 MDL)准则.MDL 的基本思想是为要随机传送的消息设计编码时,最感兴趣的编码为最简短的编码,即为了传输信息所需的最小传输位数.对获取的概念集合和文本聚类,假设集合 $\boldsymbol{W},\boldsymbol{D},\boldsymbol{C}$ 和 $\boldsymbol{T}$ 对信息接收者和发送者是透明的,只需要传输词与概念的映射关系、文本与主题的映射关系、主题与概念的映射关系以及概念个数,则当概念个数为 j 时,描述长度 MDL^j 包含 4 个部分:词与概念关系描述长度 L_C,$L_C=-\sum_{w\in W}\sum_{c\in C}P(c\mid w)\log P(c\mid w)$.文本与主题关系描述长度 L_T,$L_T=-\sum_{d\in D}\sum_{t\in T}P(t\mid d)\log P(t\mid d)$.主题与概念关系描述长度 L_{CT},$L_{CT}=-\sum_{c\in C}\sum_{t\in T}P(c,t)\log P(c,t)$.概念个数描述长度 L_{CC},也就是描述整数 j 所需的位数,$L_{CC}=2\log k_c+1$.那么选取 $k_c=\arg\min_j\{MDL^j\mid j=1,...,m\}$.

因为 m 一般很大,如果要选取全局最优 k_c,需要很长的运行时间.实际上,通过退火算法,最终获得的是一个概念层次树,也就是说,概念有其上位概念和下位概念,一个概念对其下位概念来说表达了更一般的概念,那么,本文尽可能地选取泛化的概念,即当 $MDL^{j-1}>MDL^j<MDL^{j+1}$ 时,选取 j 作为 k_c 的估计,尽管此时 MDL^j 只是局部最优解,但如第 4 节中的实验表明,经过适当泛化的概念可以取得很好的聚类结果.

3.3 算法设计

具体的概念获取与文本聚类算法见算法 1,其中步骤 5)、步骤 6)为概念 c'构造摄动向量 c_l'和 c_r',如果β达到临界值,则 c_l'和 c_r'相距应足够远,从而生成新概念.这里采用对称鉴别信息(symmetric Kullback-Leibler divergence,简称 SKL)来判断 c_l'和 c_r'的相似性,δ为分裂参数,本文设为 $1/\beta$,这主要是因为在算法初期概念较少,概念之间的距离应该比较远,而随着概念的增多,其边界也逐渐变得模糊.另外,有两点值得注意:

(1) 初始化策略.根据定理 3,由于算法 1 中在当前β值下求解 F 最小值时收敛于局部最优(步骤 7)、步骤 8)),初始化方法对算法影响非常大.步骤 1)中 $P^0(t|d)$的选择,一般常用随机初始化方法,然而本文通过实验发现,

这种方法并不理想.本文采用一种利用预聚类结果来初始化 $P(t|d)$的方法,其基本思想是利用现有的基于词特征的聚类算法将文本集合 $\boldsymbol{D}$ 聚成 k_t 个簇,并根据文本与簇的相似程度(需要标准化)来初始化 $P(t|d)$.本文在预聚类中采用了基于概率模型的划分算法[1].在步骤 2)中,由于此时 k_c 为 1,则 $P^0(c|w)=1,\forall w\in W$.

(2) 分裂策略.由于在高维空间中"胜者为王(winner-take-all)"行为的存在,算法 1 非常容易产生偏斜的结果.为了获得平衡的结果并提高聚类效率,分裂策略主要包含两部分:一是为了获得更加平衡而不是偏斜的结果,选择包含词最多的概念 c'分裂(步骤 4)).另一个是β值的调整.β值的调整(步骤 9)~步骤 11))对算法 1 的效率影响很大.如果β值增长越慢,则分裂的速度越慢,导致算法效率降低,但增长过快,又可能跨过临界值,导致质量降低.本文采用动态调整方法,$\beta^{j+1}=(1+\alpha 2^{\Delta})\beta^j$,如果自产生分裂后,$\Delta$次迭代没有产生分裂,则$\beta$值以 $1+\alpha 2^{\Delta}$倍指数增长,如果分裂,则以 $1+\alpha$倍指数增长.

算法 1. ExtractConcept&TextClustering (ECTC).

输入:联合概率分布 $P(w,d),w\in \boldsymbol{W},d\in \boldsymbol{D}$;主题个数 k_t;β值调整参数α;收敛参数ε.

输出:$P(t|d),t\in \boldsymbol{T},d\in \boldsymbol{D}$;$P(c|w),c\in \boldsymbol{C},w\in \boldsymbol{W}$ 和概念个数 k_c.

1) 初始化 $P^0(t|d),\forall t\in \boldsymbol{T},d\in \boldsymbol{D}$;$|\boldsymbol{T}|=k_t$;

2) 初始化 $P^0(c|w),\forall c\in \boldsymbol{C}^0,w\in \boldsymbol{W}$;$|\boldsymbol{C}^0|=k_c=1$;

3) While ($MDL^j>MDL^{j+1}$){

4) Let $c'=\arg\max_{c}\{\sum_{w\in W}P^j(c\mid w),c\in C^j\}$)

5) 对 c'构建概念 c_l'和 c_r',$\boldsymbol{C}^j=(\boldsymbol{C}^j-c')\cup\{c_r',c_r'\}$;$k_c=k_c+1$;

6) $\forall w\in \boldsymbol{W}$,$P^j(c_l'|w)=P^j(c'|w)(0.5+r(c',w))$和 $P^j(c_r'|w)=P^j(c'|w)(0.5-r(c',w))$; //$r(c',w)$为 0~1 的随机数,并保证 $\sum_{c\in C^j}P^j(c\mid w)=1$

7) While($F^i-F^{i+1}>\varepsilon$){

8) 计算式(14)~式(20);计算 F^{i+1};i++;}

9) 如果 $SKL(c_l',c_r')<\delta$则{$\boldsymbol{C}^j=(\boldsymbol{C}^j-\{c_r',c_r'\})\cup c'$;$k_c=k_c-1$;$\Delta$++;}

10) 否则{Δ=0;计算 MDL^{j+1};}

11) $\beta^{j+1}=(1+\alpha 2^{\Delta})\beta^j$; j++;}

12) 返回 $P(t|d)$;$P(c|w)$;k_c

4 性能分析

4.1 数据集和实验方法

本文采用 20NG(http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html)和北大燕穹的中文网页集作为测试集(http://net.pku.edu.cn/~yanqiong).本文从两个测试集中分别构造了数据集 NG4C200, TALK4C4000 和 BG2C200,BG3C600.中文分词采用了东北大学自然语言理解实验室的分词软件 CipSegSDK.英文 STEM 处理采用了 PORTER 算法.停用词处理采用人工字典方法.实验主要包括概念选取方法验证、初始化方法的比较以及文本聚类方法的比较.基于概率模型的文本聚类方法 ESPClust[1],这是一种基于词空间的聚类算法,较其他基于词空间的聚类算法有很大优势.基于信息论的双向聚类算法 Coclust[13],这是一种基于词聚类的硬聚类方法,文献[13]中的实验表明,其聚类结果好于其他基于词聚类的方法.本文选择 ESPClust 和 Coclust 作为比较算法.由于 ESPClust 和 Coclust 均为硬聚类方法,本文采用 Macro-$F1$ 作为文本聚类质量的评价指标.而 ECTC 算法是一种软聚类方法,在计算 Macro-$F1$ 时,对$\forall d\in \boldsymbol{D}$,将其划分到具有最大 $P(t|d)$值的主题中.

4.2 性能分析

首先验证本文提出的概念选取方法的有效性.图 1 为 ECTC 在 4 个数据集合上,随着概念的分裂,MDL 值的变化曲线及其相对应的文本聚类结果曲线,为了便于比较,对 MDL 值取对数.从图 1 中可以看出,当 MDL 值满

足条件 $MDL^{j-1}>MDL^{j}<MDL^{j+1}$ 时,相应的 Macro-$F1$ 值基本处于最大值,见表 2.由于篇幅所限,本文只给出了在 BG2C200 数据集上部分获取的概念层次结构,见表 1,第 1 行为概念标识,其中“\0\0\1”代表概念“\0\0\1”是由概念“\0\0”分裂而来的,表中的每一列为相对应的概念中 $P(w|c)$最大的 20 个词.第 2~4 列的概念是 ECTC 最终选取的概念,可以看出每个概念中的词都有很强的语义关系.

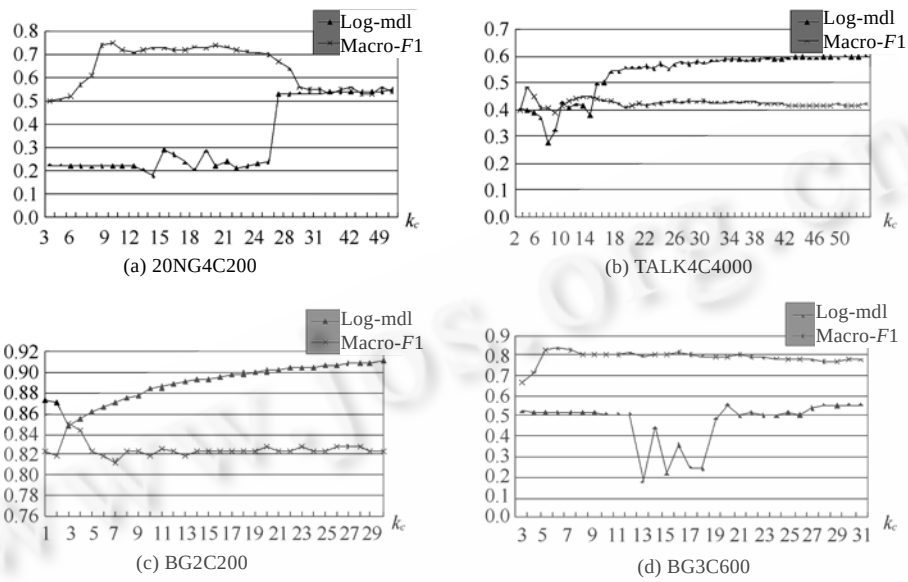


Fig.1 MDL value and the corresponding Macro-F1 of ECTC algorithm for four datasets

图 1 ECTC 在 4 个数据集上的 MDL 值与相应的 Macro-F1 值

Table 1 Illustration of partial concept hierarchy in BG2C200 dataset

表 1 在 BG2C200 数据集上的部分概念层次

\0\0	\0\0\0	\0\0\1	\0\1	\0\1\0	\0\1\1
作者	作者	服务	做生意	服务	做生意
作用	作用	形象	周年	新闻	周年
自身	自身	图文	中国	遭到	中国
智力	智力	活动	邮政	希望	邮政
指导	指导	发展	依然	文娱	依然
张纪中	张纪中	画面	演唱会	痛苦	演唱会
语言	语言	献血	巡回	时代	巡回
幼儿	幼儿	涂鸦	宣传画	泪水	宣传画
影响	影响	结果	信息	家庭	信息
引导	引导	绘画	小平	改行	小平
依赖	依赖	成人	香港	反抗	香港
医院	医院	场面	喜爱	扮演	喜爱
要素	要素	拜佛	文章	总经理	文章
要求	要求	注重	图象	主页	图象
眼睛	眼睛	舞台	深圳市	指标	深圳市
学习	学习	挑战	日前	要闻	日前
形式	形式	能否	群众	竖立	群众
新版	新版	联系	平方米	收入	平方米
心神	心神	京剧	女性	市中心	女性
小说	小说	江湖	面料	市民	面料

下面对算法的初始化方法进行比较.根据定理 3,ECTC 算法在给定的β值下只能获得局部最优解,那么如第 3 节所述,初始化方法对 ECTC 的影响很大.表 2 为 ECTC 在随机初始化与预聚类初始化两种方法下的聚类结果比较,其中 k_c 列为选取的概念个数,Macro-F1 列为在选取概念个数下 Macro-$F1$ 的值,Max 列为最大 Macro-$F1$ 值.随机初始化方法中,ECTC 运行了 5 次,并选取聚类质量最好的作为最终结果.可以看出,预聚类初始化方法在选取的概念个数上的 Macro-$F1$ 较随机初始化方法平均提高了 60.8%,最大 Macro-$F1$ 平均提高了 36.8%,并且预聚类初始化方法选取的概念个数更少,在选取的概念个数上,Macro-$F1$ 更接近于最大 Macro-$F1$.最后是 ECTC 与 ESPClust 和 Coclust 的聚类结果比较,如图 2 所示.可以看出,ECTC 的聚类质量最好,较 ESPClust 提高了 38.6%,较 Coclust 提高了 23.2%.

Table 2 Comparison of clustering under different initial methods

表 2 不同初始化方法下文本聚类结果比较

Dataset	Random			Pre-clustering			Comparison	
	k_c	Macro-$F1$	Max	k_c	Macro-$F1$	Max	Macro-$F1$ (%)	Max (%)
20NG4C200	4	0.41	0.54	13	0.72	0.74	+75.6	+37.0
TALK4C4000	10	0.25	0.25	6	0.4	0.48	+60.0	+92.0
BG2C200	3	0.45	0.79	3	0.85	0.85	+88.9	+7.6
BG3C600	16	0.69	0.75	7	0.82	0.83	+18.8	+10.7

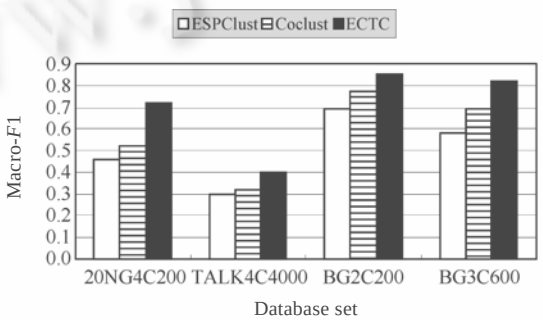


Fig.2 Comparison of Macro-$F1$ among ESPClust, Coclust and ECTC

图 2 ESPClust,Coclust 和 ECTC 的 Macro-$F1$ 比较

5 结 论

本文提出了一种基于信息论的潜在概念获取与文本聚类方法,通过实验表明,该方法要好于基于词空间的文本聚类方法以及目前基于潜在概念的文本双向聚类方法.本文的主要贡献在于:(1) 通过潜在概念变量和主题变量的引入,以及词、潜在概念、文本和主题之间关系的概率表示,反映了词与潜在概念、文本与主题和潜在概念与主题之间的模糊关系.(2) 根据信息论中熵压缩编码理论,定义了求解潜在概念和文本聚类的全局目标函数,并给出一种类似于确定性退火算法的求解算法 ECTC,用以获得概念层次树以及在不同层次概念上的文本聚类结果,是一种双向软聚类方法.(3) 提出了一种基于最短描述长度原则的概念选择方法,用以最终确定所获取的概念个数和对应的文本聚类结果.尽管该方法只是得到 MDL 的局部最优解,但实验结果表明,更为泛化的概念可以取得很好的聚类结果,并且可以获得更低维的概念空间.

References:

[1] Li XG, Yu G, Wang DL. MMPClust: A skew prevention algorithm for model-based document clustering. In: Zhou LZ, ed. Proc. of the 10th Int'l Conf. on Database Systems for Advanced Applications. Beijing: Springer-Verlag, 2005. 536−547.

[2] Deerwester S, Dumais ST, Furnas GW. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990,41(6):391−407.

[3] Hofmann T. Probabilistic latent semantic indexing. In: Proc. of the 22nd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Berkley: ACM Press, 1999. 50−57.

[4] Gong XJ, SHI ZZ. Semi-Supervised Web mining based on bayes latent semantic model. Journal of Software, 2002,13(8): 1508−1514 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/13/1508.pdf

[5] Karypis G, Han EH. Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization. In: Proc. of the 2000 ACM CIKM Int'l Conf. on Information and Knowledge Management. McLean: ACM Press, 2000. 12−19.

[6] Aggarwal CC, Yu PS. On effective conceptual indexing and similarity search in text data. In: Proc. of the 2001 IEEE Int'l Conf. on Data Mining. San Jose: IEEE Computer Society, 2001. 3−10.

[7] Baker LD, McCallum AK. Distributional clustering of words for text classification. In: Proc. of the 21st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Melbourne: ACM Press, 1998. 96−103.

[8] Dhilon IS. Co-Clustering documents and words using bipartite spectral graph partitioning. In: Proc. of the 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Francisco: ACM Press, 2001. 269−274.

[9] Tishby N, Pereira FC, Bialek W. The information bottleneck method. In: Proc. of the 37th Annual Allerton Conf. on Communcation, Control and Computing. 1999. 368−377.

[10] Slonim N, Tishby N. Document clustering using word clusters via the information bottleneck. In: Belkin NJ, *et al.*, eds. Proc. of the 23rd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Athens: ACM Press, 2000. 208−215.

[11] Pereira F, Tishby N, Lee L. Distributional clustering of English words. In: Proc. of the 31st Annual Meeting of the Association for Computational Linguistics. Columbus: Morgan Kaufmann Publishers, 1993. 183−190.

[12] Friedman N, Mosenzon O, Slonim N, Tishby N. Multivariate information bottleneck. In: Breese JS, Koller D, eds. Proc. of the 17th Conf. on Uncertainty in Artificial Intelligence. Seattle: Morgan Kaufmann Publishers, 2001. 152−161.

[13] Dhillon IS, Mallela S, Modha DS. Information theoretic co-clustering. In: Getoor L, Senator TE, Domingos P, Faloutsos C, eds. Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Washington: ACM, 2003. 89−98.

[14] Rose K. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. Proc. of the IEEE, 1998,86(11):2210−2239.

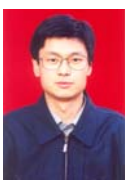
附中文参考文献:

[4] 宫秀军,史忠植.基于 Bayes 潜在语义模型的半监督 Web 挖掘.软件学报,2002,13(8):1508−1514. http://www.jos.org.cn/1000-9825/13/1508.pdf

李晓光(1973－),男,辽宁沈阳人,博士,副教授,主要研究领域为数据挖掘,信息检索,流数据分析.

王大玲(1962－),女,博士,教授,CCF 高级会员,主要研究领域为数据挖掘,Web 挖掘.

于 戈(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库理论与技术.



鲍玉斌(1968－),男,博士,副教授,CCF 高级会员,主要研究领域为数据仓库,OLAP.