模糊集间的语义关联度及其应用*

何新贵

(北京系统工程研究所,北京 100101)

摘要 本文提出了一种描述模糊集间近似程度的语义关联度概念.它不仅与模糊集论域元素的隶属度有关,而且考虑了论域元素间的语义近似程度.因此它比过去模糊数学中定义的各种距离和贴近度等概念更加精细地刻划了模糊集间的相似性,从而在人工智能和其它领域中可有广泛应用,文中列举了它在情报检索和正文分类等方面的应用实例.此外,论文还给了两种近似地计算语义关联度的算法.

关键词 语义关联度*,模糊集.

1 问题的提出

为了使现有的计算机能处理与自然语言有关的各种问题,让机器能理解自然语言自然是一条求之不得的途径.但由于自然语言的复杂性,这条途径困难很大,可是客观上又有许多与自然语言有关的问题等待我们去解决.

对用自然语言书写的文献资料的检索是一个古典问题,但应该说至今并没有完满解决.要在浩瀚的文海中找到你所需要的材料,首先面临一个如何把你的检索意图告诉计算机的问题.传统上有"按标题检索","按作者检索"和"按关键字检索"等方式.但当这些用于检索的信息不知道或知道得不清楚(或不确切)时这些方法就失灵了.这就提出了是否有更好检索方法的问题.

此外,用自然语言书写的正文的分类问题也是经常遇到的另一个问题,其中图书资料的分类就是一例.这些事情如果全由人来做,由于人的主观随意性很大,不同的人往往有不同的处理原则和标准,很难对此做到一致而客观的解决.可见,采用计算机来辅助处理这类问题也势在必行.

由于自然语言理解的彻底解决似乎在短时期内尚无希望.因此开拓一条现实可行的途径,尽管可能解决得不太理想,或许只能作为人的辅助工具,仍然是很有意义的.为此,我们首先要来引进一个可作为正文分类依据的关于模糊集的语义关联度的概念.并指出这种语义关联度的各种可能的应用.

* 本文1991—04—25收到,1992—02—10定稿

作者何新贵,研究员,主要研究领域为数据库,知识处理,模糊处理的理论与技术研究.

本文通讯联系人:何新贵,北京 100101,北京系统工程研究所

2 模糊集间的语义关联度

在过去,为了实现在一定程度上的按正文语义进行检索,采用了对每篇文献或情报资料抽取几个最相关的"关键字",以便在一定程度上用这个"关键字"的集合来近似地表示原文的语义.它或可作为检索的条件,或可作为分类的依据.在此我们想将此方法进行推广,就是采用"模糊关键字集"来取代一般关键字集,以便用它更精确地描述原文的语义.设 $K=\{p_1/k_1,p_2/k_2,\cdots,p_n/k_n\}$,是一个模糊集合,其中 $k_1,k_2,\cdots,k_n$ 是 n 个关键字; $p_1,p_2,\cdots,p_n$ 是 n 个取值于[0,1]的相应关键字的隶属度.隶属度在不同的场合可用来表示关键字在原文中的"重要性"、"代表性"、"出现频率"以及更抽象的其它度量含义等.

在[1]中,作者曾提出过语义距离的概念,可用它来度量两个模糊数或更复杂的模糊对象之间的相近程度.语义距离已被成功地应用在模糊数据库中进行模糊检索.为了更好地解决上述情报检索和正文分类之类的应用问题,现在我们来引进模糊集之间的语义关联度的概念.

首先来定义一般集合间的语义关联度.设

$$A=\{a_1,a_2,\cdots,a_n\},B=\{b_1,b_2,\cdots,b_m\},m\geqslant n \text{ 且 } |m-n|\ll \min(m,n)$$

为两个有限集.为了使 A 的每元素唯一地对应 B 的某元素,显然共有 P_m^n 种对应方式:

$$a_i\longleftrightarrow b_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,P_m^n)$$

其中 b_{ij} 是 B 的不同的元素.假设我们已在 A 的元素 a 与 B 元素 b 之间定义了语义关联度 SR(a,b),它是 0 与 1 之间的一实数,0 表示语义无关,1 表示语义相等,且设 SR(a,b)=SR(b,a).

定义 1. 集合 A 和集合 B 之间的语义关联度定义为

$$SR(A,B)=\max_{1\leqslant j\leqslant P_m^n}\left(\frac{1}{n}\sum_{i=1}^{n}(SR(ai,b_{ij}))^p\right)^{\frac{1}{p}} \quad (\text{设 } m\geqslant n),\text{并令 } SR(B,A)=SR(A,B),$$

其中 $p\geqslant 1$,特别当 p=1,2 和∞时分别为

$$SR_1(A,B)=\max_{1\leqslant j\leqslant P_m^n}\frac{1}{n}\sum_{i=1}^{n}SR(a_i,b_{ij})$$

$$SR_2(A,B)=\max_{1\leqslant j\leqslant P_m^n}\sqrt{\frac{1}{n}\sum_{i=1}^{n}SR(a_i,b_{ij})^2}$$

和

$$SR_\infty(A,B)=\max_{1\leqslant j\leqslant P_m^n}\max_{1\leqslant i\leqslant n}SR(a_i,b_{ij})$$

易见,当 $A\subseteq B$ 时 SR(A,B)最大,表示 A 与 B 最为相关;当且仅当集合 A 和 B 的两两元素间语义关联度全都为零时,SR(A,B)才等于零,表示 A 与 B 完全无关.且有 $0\leqslant SR(A,B)\leqslant 1$.

现在我们进一步来把上定义推广到模糊集上.

定义 2. 设

$$\underset{\sim}{A}=\{p_1/a_1,p_2/a_2,\cdots,p_n/a_n\}$$

$$\underset{\sim}{B}=\{q_1/b_1,q_2/b_2,\cdots,q_m/b_m\},\quad m\geqslant n,\text{且}|m-n|\ll\min(m,n)$$

为两个模糊集.类似地,可建立 p_m^n 种对应关系:

$$a_i \longleftrightarrow b_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,p_m^n)$$

其中 $b_{ij}(j=1,2,\cdots,p_m^n)$ 是 $\underset{\sim}{B}$ 的论域中的不同的元素. 假设在 $\underset{\sim}{A}$ 和 $\underset{\sim}{B}$ 的论域的元素 a 与 b 之间已经定义了语义关联度 SR(a,b)，其含义同前. 模糊集 $\underset{\sim}{A}$ 和 $\underset{\sim}{B}$ 之间的语义关联度定义为：

$$SR(\underset{\sim}{A},\underset{\sim}{B})=\max_{1\leqslant j\leqslant p_m^n}\{\frac{1}{n}\sum_{i=1}^{n}[(1-|p_i-q_{ij}|)*SR(a_i,b_{ij})]^p\}^{\frac{1}{p}},(m\geqslant n),$$

并令 $SR(\underset{\sim}{B},\underset{\sim}{A})=SR(\underset{\sim}{A},\underset{\sim}{B})$

其中 p_i 与 q_{ij} 分别为 a_i 与 b_{ij} 的隶属度；* 是某种二元运算，例如乘法和求极小等；$p\geqslant 1$，特别当 p=1，2，和∞时分别为

$$SR_1(\underset{\sim}{A},\underset{\sim}{B})=\max_{1\leqslant j\leqslant p_m^n}\frac{1}{n}\sum_{i=1}^{n}[(1-|p_i-q_{ij}|)*SR(a_i,b_{ij})]$$

$$SR_2(\underset{\sim}{A},\underset{\sim}{B})=\max_{1\leqslant j\leqslant p_m^n}\sqrt{\frac{1}{n}\sum_{i=1}^{n}[(1-|p_i-q_{ij}|)*SR(a_i,b_{ij})]^2}$$

和

$$SR_\infty(\underset{\sim}{A},\underset{\sim}{B})=\max_{1\leqslant j\leqslant p_m^n}\max_{1\leqslant i\leqslant n}\{(1-|p_i-q_{ij}|)*SR(a_i,b_{ij})\}$$

易见，$SR(\underset{\sim}{A},\underset{\sim}{A})=1$，这即说任一模糊集与其自身的语义关联度最大. 且有 $0\leqslant SR(\underset{\sim}{A},\underset{\sim}{B})\leqslant 1$. 称 $SR'(\underset{\sim}{A},\underset{\sim}{B})=SR'(\underset{\sim}{B},\underset{\sim}{A})=\max_{1\leqslant j\leqslant p_m^n}\{\sum_{i=1}^{n}[(1-|p_i-q_{ij}|)*SR(a_i,b_{ij})]^p\}^{\frac{1}{p}}$

为 $\underset{\sim}{A}$ 与 $\underset{\sim}{B}$ 之间非规范化的语义关联度. 这种定义方式反映了当在 $\underset{\sim}{A}$ 中增加“元素——隶属度”对（其它元素的隶属度不变）时，$SR'(\underset{\sim}{A},\underset{\sim}{B})$ 的值只可能增加或不增，这在某些应用场合是合理的，但本文将只讨论上述已经规范化的关联度 $SR(\underset{\sim}{A},\underset{\sim}{B})$. 若 $\underset{\sim}{A}$ 和 $\underset{\sim}{B}$ 论域元素间的一种对应关系，$a_i\longleftrightarrow b_{ij}$，(i=1,2,…,n) 满足：

$$SR(\underset{\sim}{A},\underset{\sim}{B})=\{\frac{1}{n}\sum_{j=1}^{n}[(1-1p_i-q_{ij}|)*SR(a_i,b_{ij})]^p\}^{\frac{1}{p}},$$

则称该对应关系是最佳对应.

由定义可见，两个模糊集间的语义关联度，不但依赖于这两个模糊集论域的元素间的语义关联度 $SR(a_i,b_{ij})$，而且与相应论域元素 a_i 与 b_{ij} 的隶属度 p_i 与 q_{ij} 有关. 这说明，能在对应的论域中找到语义关联度大的元素对还不足以保证两个模糊集的语义关联性强，为此还必须使相互对应的元素的隶属度也相近才行. 这样显然比非模糊集间的语义关联度的描述更精细了. 此外，当在 $\underset{\sim}{A}$ 中增加“元素——隶属度”对（其中其它元素的隶属度不变）时，若能在 $\underset{\sim}{B}$ 中找到与它的关联度大的“元素——隶属度“对时，就可能增加 $\underset{\sim}{A}$ 和 $\underset{\sim}{B}$ 之间的语义关联度，否则若在 $\underset{\sim}{B}$ 中找不到与之关联度大的“元素——隶属度”对时，反而可能减少 $\underset{\sim}{A}$ 和 $\underset{\sim}{B}$ 之间的语义关联度，这是十分符合人们的直觉的. 以上的讨论说明了上述语义关联度的定义是合理的，是符合实际的. 以前关于模糊集间的各种距离和贴近度[2]等的定义中都没有反映论域元素间的语义关联性. 这对文献资料检索和正文分类等应用是不合适的，而语义关联度的描述正好满足了这种要求，它已在一些实际应用中得到了满意的结果. 下面我们来举例说明语义关联度概念的几种可能的应用.

3　模糊关键字集的抽取

为了把模糊集间的语义关联度概念应用于资料检索或正文分类，首先必须解决如何从

资料原文或正文中抽取一个可以近似描述其语义的模糊关键字集的问题.这就是说,不但要选择相关的关键字,而且要确定其相应的隶属度(可表示重要度,频度和代表性等含意),构成一个模糊关键字集合.

如何根据正文的语义抽取这种可近似表示正文语义的关键字集是一个需要较高智能的问题.严格讲除了要求理解正文的含义之外,尚需有总结概括的能力乃至有较深的领域知识,才能较好地解决这问题.直至今日,这仍是难以用现有计算机来实现的.在此我们建议采用受限自然语言理解再结合专家系统[3]的办法来较现实地处理它.因此必须与语言学家们结合把人类在抽取正文关键字时所遵循的原则总结出来.例如,可能的抽取原则可以包括:

* 正文中的诸如前置词、冠词、代词等词类一般不在被选择之列.形容词与副词若被选中,必须与其修饰的词结合在一起,作为一个关键字.

* 若在原正文中已被选为关键字(如果有的话),则也选中它,并给予隶属度1.

* 在标题和摘要(如果有的话)中的词有最大的可能性被选中,并给予较高的隶属度.

* 根据受限自然语言理解的途径,找出正文中的一些"关键句",即那些包含诸如"关键在于…","旨在…","主要目的(标)是…"等的句子.在"关键句"中的词有相当大的可能性被选中,而且给予较大的隶属度.

* 在引言和结论段中的词有较大的可能性被选中,并给予一定的隶属度.

* 在段首或段尾出现的词有较大的可能性被选中,并给予一定的隶属度.

* 要重视选择出现频度高的词,并随着频度的增高逐次增加其隶属度.

* 隶属度叠加原则,即若一个词同时处于上述多种地位,则其隶属度以某种方式叠加.

* 同义词、近义词或转义词出现时,根据其间的语义关联度大小作为某关键词的一次或部分出现,统计在出现频数中.

* 应使最终选出的关键字两两之间语义关联度很小,即应使选出的模糊关键字集的论域元素之间语义上尽量无关.

一般,采用上手段获得的"模糊关键字集"还应进行"λ—滤波操作"[4],即把该模糊关键字集中隶属度小于λ(0<λ≤1)的关键字滤去掉.这样就可把不够重要的关键字略掉,而最终得到一可以近似描述原正文语义的一个"模糊关键字集".在此需要指出,不同的场合选用多大的λ值来进行滤波要根据实际情况而定,不能一概而论.当然,为了比较两篇正文的关联程度,应该采用相同的产生"模糊关键字集"的方法和相同的滤波基数λ.

4 文献资料的检索

模糊集间的语义关联度的第一个应用是正文(包括文献资料和文件等)的检索.为此存放在库中的正文应首先采用上节所述的方法抽取模糊关键字集,并在此模糊关键字集与原正文之间建立指针联系,以便从模糊关键字集可以沿指针方便地找到其相应的原正文.一旦建立了这种正文库以后,用户就可用一个"模糊关键字集"$\underset{\sim}{K}$来进行检索了.即计算$\underset{\sim}{K}$与库中各正文的模糊关键字集$\underset{\sim}{K}_i$之间的语义关联度$SR(\underset{\sim}{K},\underset{\sim}{K}_i)$,若对某量λ:0<λ<1有$SR(\underset{\sim}{K},\underset{\sim}{K}_i)>\lambda$,则称$\underset{\sim}{K}_i$所代表的正文$T_i$和$\underset{\sim}{K}$是λ—关联的.不妨就取$T=\{T_i|\underset{\sim}{K}_i$是$T_i$的关键字集,且$SR(\underset{\sim}{K},\underset{\sim}{K}_i)>\lambda\}$作为用$\underset{\sim}{K}$从库中检索的结果.这里可选取不同的λ值来控制检索的

“精度”. 显然，λ 越大，能从库中检索的结果越少，而且所得的结果正文的语义越能被检索关键字集 $\underset{\sim}{K}$ 所近似地表示. 一般在实际应用中，要求尽量使 $\underset{\sim}{K}$ 与 $\underset{\sim}{K}_i$ 的论域元素的个数(即关键字数)相等.

5　正文分类

正文分类是模糊集语义关联度的又一个应用，设 $\underset{\sim}{C}_1,\underset{\sim}{C}_2,\cdots,\underset{\sim}{C}_n$ 为 n 个模糊关键字集，(它们各自包含的关键字的个数相同)称为分类基集. 它们分别代表着 n 个正文类. 今有一个正文 T，要确定 T 最好应该分到哪类去. 为此可按下列步骤处理：

1. 按第 3 节中所述方法建立正文 T 的模糊关键字集 $\underset{\sim}{T}$，其关键字个数应与 $\underset{\sim}{C}_i$ 的尽量相同.

2. 分别计算　　$SR(\underset{\sim}{T},\underset{\sim}{C}_i),\quad i=1,2,\cdots,n$

3. 选取　　$J=\{j\mid SR(\underset{\sim}{T},\underset{\sim}{C}_j)=\max\limits_{1\leqslant i\leqslant n} SR(\underset{\sim}{T},\underset{\sim}{C}_i)\}$,

表示正文 T 应被分类在第 j 类正文中，这里 $j\in J$. 由于 J 中可能有多个元素，所以 T 可能同时被分在多个类中. 如果必要还可采用进一步的分类原则将 T 确定唯一的分类. 相反，若要更糊一点的分类，则不妨把上述选取 J 的第 3 步改为：对一个小量 $\varepsilon>0$，选取

$$J=\{j\mid |SR(\underset{\sim}{T},\underset{\sim}{C}_j)-\max_{1\leqslant i\leqslant n}SR(\underset{\sim}{T},\underset{\sim}{C}_i)|\leqslant\varepsilon\}$$

为了使上述方法真正实用化，问题可能在于如何简便地计算两个模糊集间的语义关联度. 从定义中可见，为计算 $SR(\underset{\sim}{A},\underset{\sim}{B})$ 需要 p_m^n 次求和或取极大等操作，当 m，n 较大时是难以接受的. 下面我们来给出几种近似地计算语义关联度的方法.

6　语义关联度的近似计算

首先，两个模糊集论域元素间的语义关联度要根据论域的具体情况，由具体的领域专家们来给出，例如在正文检索和分类中，论域是关键字的集合，应由语言学家们来给出一本“同义与近义词关联度字典”，给出各关键字间两两的关联程度. 一旦有了这种字典以后，在计算两个模糊集间的语义关联度时，就可根据需要去查相应的关联度字典. 有两种途径可用来解决语义关联度的计算量很大这个问题，一种途径是采用迭代算法逐次逼近精确值，另一途径是给出一些计算其近似值的方法. 关于前一种途径我们将另文介绍. 下面我们来给出采用后一途径的两种近似算法：

设

$$\underset{\sim}{A}=\{p_1/a_1,p_2/a_2,\cdots,p_n/a_n\},$$
$$\underset{\sim}{B}=\{q_1/b_1,q_2/b_2,\cdots,q_n/b_m\},m\geqslant n$$

算法 1：

1. $i=1$；　$SR(\underset{\sim}{A},\underset{\sim}{B})=0$；
2. $DB=\{b_1,b_2,\cdots,b_m\}$
3. 任取 $b_k\in\{b'_k\mid(1-|p_i-q_k|)*SR(a_i,b'_k)=\max\limits_{b_j\in DB}[(1-|p_i-q_j|)*SR(a_i,b_j)]\}$
4. $SR(\underset{\sim}{A},\underset{\sim}{B})=SR(\underset{\sim}{A},\underset{\sim}{B})+SR(a_i,b_k)*(1-|p_i-q_k|)$
5. $i=i+1$；
6. $i>n$?；　是转 8，　不是转 7；
7. $DB=DB-\{b_k\}$；转 3；

8. exit(给出结果 $SR(\underset{\sim}{A},B)$);

这种算法比较简单,但可能产生较大的误差,下面的算法 2 比算法 1 更精细,但包含着较复杂的计算.

算法 2:

1. $SR(\underset{\sim}{A},B)=0$;
2. $DA=\{a_1,a_2,\cdots,a_n\}$; $DB=\{b_1,b_2,\cdots,b_m\}$;
3. $dB=\{b_k \mid \max\limits_{i\in DA}[(1-|p_i-q_k|)*SR(a_i,b_k)=\max\limits_{b_j\in DB}\max\limits_{i\in DA}[(1-|p_i-q_j|)*SR(a_i,b_j)]],a_i\in DA\}$;
4. 对任一 $b_k\in dB$,令 $dA_k=\{a_i \mid (1-|p_i-q_k|)*SR(a_i,b_k)=\max\limits_{b_j\in DB}[(1-|p_i-q_j|)*SR(a_i,b_j))],a_j\in DA,b_k\in dB\}$;

   任选 $a_l\in\{a'_l \mid (1-|p_l-q_k|)*SR(a'_l,b_k)=\max\limits_{a_i\in dA_k}[(1-|p_i-q_k|)*SR(a_i,b_k)],b_k\in dB\}$;
5. $SR(\underset{\sim}{A},B)=SR(\underset{\sim}{A},B)+(1-|p_l-q_k|)*SR(a_l,b_k)$;

   $dB=dB-\{b_k\}$; $DB=DB-\{b_k\}$; $DA=DA-\{a_l\}$;
6. dB=空? 空则转 7;不空则转 4;
7. DA=空? 空则转 8;不空则转 3;
8. exit(给出结果 $SR(\underset{\sim}{A},B)$);

参考文献

1 何新贵. 模糊数据库中的语义距离及模糊视图. 计算机学报,1989,12(10):757—764.

2 贺仲雄. 模糊数学及其应用. 天津:天津科学技术版社,1983.

4 何新贵. 知识处理与专家系统. 北京:国防工业版社,1990.

3 何新贵. 模糊关系数据库的数据模型. 计算机学报,1989,12(2):120—126.

SEMANTIC RELATIONSHIP BETWEEN FUZZY SETS AND ITS APPLICATIONS

He Xingui

(*Beijing Institute of System Engineering*, *Beijing* 100101)

Abstract A concept of semantic relationship between fuzzy sets is presented in the paper. The concept is not only related to the membership of the fuzzy sets, but also considered releant to the similarity between elements in the domains of the fuzzy sets. So, it can be used to characterize the similarity between fuzzy sets more meticulously than it was done by the various "distance" or "nearness" previously defined in fuzzy mathematics. Therefore, it would have a wide range of applications in AI and other areas, such as information retrieval and text classification. Besides, two algorithms to approximately calculate the semantic relationship are given in the paper.

Key words Semantic relationship*, fuzzy set.