

以本体构造中文信息过滤中的需求模型

袁兴宇¹，王挺¹，周会平¹，肖君²

¹国防科技大学计算机学院，²上海远程教育集团

摘要：在信息过滤系统中，用户模板是机器可理解的用户需求表示形式，是否能准确地反映出用户的真实需求将直接影响着过滤系统的性能。在向量空间模型中，用户的模板表现为一组带权重的特征词集，但由于在这样的用户模板中缺少必要的语义信息，很难准确地反映出用户的需求。本文提出了以本体构造需求模板的方法，以本体的形式定义需求中概念间的语义关联关系，将向量空间模型中的特征向量定义为本体中的实例，通过实例间的关联路径计算特征项间的语义关联，并通过特征项间的语义关联计算出文档与模板的语义关联度。

关键词：信息过滤；本体；语义关联；用户模板

1. 引言

信息过滤就是从动态的信息数据流中查询满足用户特定需求信息的过程。这种用户的特定需求在信息过滤系统中表示为一个用户模板，对流入的信息流，根据用户定义的模板来判断某信息是否满足于用户的需求。因此，用户模板是否能真实而准确的反映出用户的需求将直接影响着过滤系统的性能，成为影响过滤性能的主要瓶颈之一。在信息检索与信息过滤中，为了使机器能更为准确地理解用户提交的查询需求和待处理的信息内容，本体、语义理解、知识推理等技术的应用变得越来越广泛。现今，在对信息的处理及理解上，应用的最多的两个通用型本体就是英文的 WordNet 和中文的“知网”（HowNet）。在对内容的处理上，[1]利用“知网”建立了真实文本的概念关系图，并在此基础上对文本的内容进行了基于理解的推理；[2]则是利用 WordNet 将文档表示为一个带有分值的概念节点集合来表示文档中的语义内容。在对用户的需求表述上，[3]、[4]提出了以语义框架构造用户需求模型的方法，利用已有的概念层次词典将用户的需求以一种语义框架的结构来表述，这种语义框架其实就可认为是一种本体结构模型。

本文以 owl 作为本体的描述语言来构造用户的需求模型，利用本体中各个节点间关联体现出在需求中各概念间的语义关系，并根据节点间关联路径，反映出概念间语义关联的强弱。本文所构造的本体属于任务本体，是针对某一具体需求而构建的本体，根据用户需求中的内容和语义倾向，将本体划分为不同区域来反映需求中的语义倾向和具体的需求。

本文下面的内容主要包括：第二节介绍本过滤系统的体系结构；第三节介绍本体模板的构造与描述；第四节为语义关联强度计算；第五节为实验设计及结果分析；第六节给出了结论以及今后的进一步工作。

2. 系统体系结构

图 1 为本过滤系统的体系结构图：

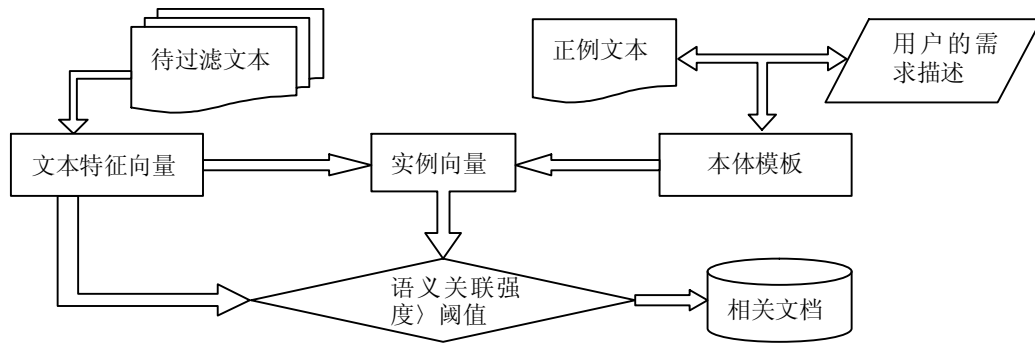


Fig.1 Architecture of filtering system

本系统主要的核心部分是本体模板的构造和语义关联强度的计算。在本体模板的构造上，从用户给出的需求描述和提供的正例文本出发，将需求中的特征项定义为本体中实例，将特征项间的语义关系定义为本体中的属性关联。在语义关联强度的计算上，将特征项与本体中相应的实例相关联，组成与特征向量相对应的实例向量。根据本体中实例间的关联关系，以实例间的关联路径作为它们间的关联强度，计算实例向量中各实例的语义关联强度。最后，计算文本向量与实例向量的内积值作为文本与本体模板的语义关联度，判断是否符合用户需求。

3. 基于本体的用户需求模型

3.1 本体模板的构造

本体所描述的不仅仅是资源的分类以及资源间关系，更重要的是能通过它们间关系推理出它们间潜在的关联关系，这对基于知识的查询与内容处理上十分重要。如例句“我骑着自行车回家。”在词典中“自行车”与“家”的义项没有任何的语义关联，但将该句定义在本体中时，“家”与“自行车”是有一定的语义关联的，这种关联由中间的节点“我”建立起来，关联的中间节点越多，说明两个词间的语义关联越弱。

对以本体来表示的需求模型来说，构造需求本体就是定义需求描述中的特征词所表述的语义以及它们间的语义关联。在定义本体的各种组件（类、属性、实例）时，所使用到的术语集就是从需求描述和正例文本中提取的特征词所组成的特征词集。因此构造本体模板时，我们主要遵循以下几点原则：

第一，在对本体中的术语选择上，主要根据特征词对主题需求的贡献程度，并不是根据特征词在向量空间上的权重。

第二，在对同义词项的处理上，将每一个类节点都定义一个“同义”的数值属性，扩充该类实例术语的同义项。

第三，在对实例术语的粒度控制上，以单个词作为实例术语，并为该实例创建了一个“限定”的数值属性，限定、明确该实例术语在本体中的语义，这点类似于词对共现的方法。

第四，在对术语间的关联的定义上，需尽可能直接、明确，不定义术语间的潜在关联，根据用户语义和需求倾向将本体中的关联以及实例划分为若千的区域，并根据用户对不同区域的关注程度不同，赋予各个区域相应的权值，来进一步的明确用户的具体需求和语义倾向。

第五，在定义本体中术语间的关联时，有一些关联对需求的语义及观点的表述贡献很强，这种关联的有无决定着两个类间的语义以及需求中观点的倾向，我们将这种关联定义为强式

关联。而将术语间内在的语义关联或不会对用户的需求造成很大影响的关联定义为弱式关联。例如，在关于伊拉克方面的新闻中，特征词“伊拉克”与“美国”之间在不同内容的新闻中，它们的语义关联也不同。在本体中对它们之间关联的不同定义，也将会反映出用户不同的需求，因此，它们之间的关联就属于强式关联，而“美国”与“布什”之间的关联就属于的是弱式关联，因为，无论它们在何种语料中出现，它们间的语义关联都是相同的。

3.2 需求的本体描述

本系统使用 OWL 作为本体模板的描述语言，并使用 protégé 作为构造本体模板的工具，根据 ontology 模型的定义[7]，本系统的需求本体表示模型定义如下：

定义 1：用户需求本体模型，将该本体模型表示记作 $O = \langle T, R, I, CI \rangle$ ；其中 T 为术语集，包括本体表示中的类术语 TC 、实例术语 TI 和关系术语 TR ； R 为关系集，声明类之间和实例之间的关系； I 为实例集，声明在本体中定义的实例； CI 为实例的声明集，用来声明类术语的实例。

定义 2：对于给定的本体 $O = \langle T, R, I, CI \rangle$ ，关系集 R 包括类属性和数值属性，类属性表示类间的关系，数值属性表示类的特性，也可将数值属性称为实例属性；将类属性定义为 $RC = TR(A, B)$ ， $A, B \in TC$ ；数值属性定义为 $RI = TR(A, B)$ ， $A \in TI, B \in \text{数值类型}$ 。

定义 3：对于给定的本体 $O = \langle T, R, I, CI \rangle$ ，实例集 $I = \{TI_j / TI_j(RI)\}$ ， $(1 \leq j \leq |TI|)$ ，其中 RI 为该实例所拥有的数值属性。

定义 4：对于给定的本体 $O = \langle T, R, I, CI \rangle$ ，实例的声明集 $CI = \{TC_j(TI_1, TI_2, \dots, TI_n)\}$ ， $(1 \leq n \leq |TI_j|)$ ， $(1 \leq j \leq |TC|)$ ，其中 TI_j 为类 TC_j 的实例。

根据 3.1 中构造本体的原则，由以上定义创建一个以“伊拉克战后重建”为需求描述的本体表示模型。具体需求描述为：“相关新闻为关于伊拉克战后重建所面临的各种困难和挑战，以及国际社会对这些困难提出的各种解决方案。与分析重建困难无关的关于伊拉克战后重建的新闻视为无关”。

由于篇幅所限，以下只给出本体模板中的小部分描述。关于伊拉克重建的本体模型表示为：

O -伊拉克 = $\langle T, R, I, CI \rangle$ ，在术语集 T 中， $TC = \{\text{人员, 援助组织, 受援组织, 方案, 困难}\}$ ； $TI = \{\text{安南, 联合国, 伊拉克, 债务, 制裁, 种族, 治安}\}$ ； $TR = \{\text{重建, 提出, 限定, 面临, 属于, 同义}\}$

关系集 R 中， $R = \{\text{重建(援助组织, 受援组织), 提出(援助组织, 方案), 面临(受援组织, 困难), 属于(人员, 援助组织), 同义(T, String), 限定(困难, String)}\}$

实例集 I 中， $I = \{\text{安南, 伊拉克, 债务(限定(方案, 免除)), 制裁(限定(方案, 取消)), 种族(限定(困难, 争斗))}\}$

实例声明集 CI 中， $CI = \{\text{人员(安南), 援助组织(联合国), 受援组织(伊拉克), 方案(债务, 制裁), 困难(种族, 治安)}\}$

根据用户的主题描述，用户对“重建所面临的困难”、“国际社会提出的解决方案”、“伊拉克重建”这三个方面较为关注，但并不是所有与伊拉克重建相关的内容均满足用户的需求，由此将本体划分为三个区域，并赋予相应的区域权重： $r_1=0.8$ ， $r_2=0.8$ ， $r_3=0.5$ ，反映出用户的需求倾向，并在 R_2 中定义强式关联属性“重建”，明确援助组织与受援组织之间的关系。图 2 为在本体模板中类的关系图结构，以及划分的不同区域。

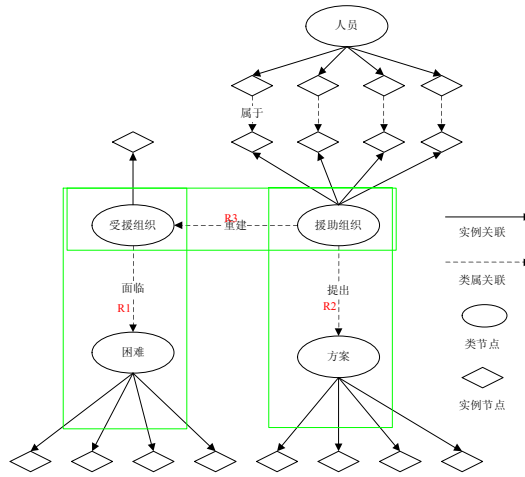


图2 伊拉克重建本体及划分的区域

Fig.2 Overview of the ontology about postwar rebuilding in Iraq

4. 语义关联计算

4.1 实例语义关联计算

本文中计算实例间语义关联的方法类似于[5]中计算实例间语义关联的方法，将本体中实例节点的语义关联强度由它们之间的路径长短来表示，越短则表明两实例的语义关联越强。将两实例节点间所经路径定义为 P ，对拥有多条路径的两节点，取其最短路径作为这两个实例节点的语义关联强度。按对本体关联划分的不同区域计算实例集 I 中各实例间的关联强度 S_p 。为避免同属相同类别的特征项大量聚集而导致的语义偏移情况的发生，同属相同类的实例之间的关联将不做计算，记为 $S_p=0$ 。

$$S_p = \frac{r_i + 1}{|c|}, P \subseteq R_i, c \in P \quad (1)$$

其中 r_i 为各个区域的权重， P 为两个实例节点间的最短路径。 c 为属于该路径的所有组件（不包括路径 P 的起始节点和终止节点）[5]，并且路径 P 上的所有组件 c 均属于区域 R_i 。当一个类属于某个区域时，那么与该类相关联的属性也被认定为属于该区域。当两个实例间的路径经过多个区域时，取 $r_i=0$ 。

例如，在伊拉克的本体中，“困难”类的实例“债务”与“援助组织”类的实例“美国”之间路径 P 为：债务—实例关联—困难—面临—受援组织—重建—援助组织—实例关联—美国， P 经过区域 $R1$ 和区域 $R3$ ，则两实例间的关联强度为：

$$S_p = \frac{1}{7} = 0.143$$

4.2 文档语义关联度计算

在对文档的表示上，我们仍然采用特征向量的表示形式，文档与本体模板的语义相关度就是特征项向量所承载的语义关联强度。将每一个特征项与模板中的实例作关联匹配，由此得到由相关实例所组成的向量 $I = (C_1I_1, \dots, C_mI_m)$ 。特征项如果没有在本体中定义，则它不予任何实例关联。如果已定义，那么特征项与相应实例的匹配以及它们之关联关系，需从以下三个方面予以考虑：

1. 强式关联。确定特征向量中是否包含本体中定义的强式关联属性词，如果某个强式关联词不存在，那么在本体中所有经该强式关联的路径均被认为不能连通，如果经该强式关联的两个实例节点还有其他连通路路径的话，则它们间的语义关联需重新计算。
2. 无“限定”属性值。如果该实例无“限定”属性值，则直接将该特征项与相应的实例相关联。
3. 有“限定”属性值。如果该实例含有“限定”属性，则需判断该特征词在文档中出现的位置，长度为3的窗口内是否有“限定”属性的值，如果没有，则该特征项也不予任何实例相关联。例如，特征项“债务”，则判断“债务”在文中左右各3个词内是否有“限定”属性值“巨额”、“沉重”等出现，如果有，则将“困难”类的实例“债务”与该特征项相关联。

由公式（1）计算实例向量 I 中实例间的语义关联强度 S_p ，如果特征项不予任何实例相关联，那么该项对应的 $S_p=0$ 。将某实例与其余各实例的语义关联强度的算术平均值作为该实例在实例向量 I 中的语义关联强度，由此构造一个实例的语义关联强度向量 $SI=(S_1, S_2, \dots, S_n)$ 。将文本特征向量和实例的语义关联强度向量作内积，设文本向量为 $D = (d_1, d_2, \dots, d_n)$ ，则 D 与 SI 间的内积表达式如下：

$$\text{RelationStrength} = \sum_{i=1}^n d_i \cdot S_i \quad (2)$$

在公式（2）中， $d_i \cdot S_i$ 表示的就是特征项 d_i 对文本 D 的语义贡献（这里的语义贡献是相对于对需求的语义贡献来说的），值越大说明该特征词在文本中越重要，表达的语义也越强。而计算得出的内积值表示的就是该文本与本体模板语义相关联的程度， RelationStrength 值越大，说明该文本表述的语义与本体模板中定义的语义相关性越强。

5. 实验设计与结果分析

5.1 测试语料与评测方法

对系统进行评测的语料有两部分：第一部分是从各大权威新闻网站中人工收集的关于伊拉克方面的新闻语料 622 篇，其中，与主题相关的页面 18 篇，关于伊拉克重建但却不与主题相关的页面 15 篇，其余为关于伊拉克战后各个方面的新闻语料；第二部分是利用 Teleport 工具，从人民网中关于伊拉克战争专题中收集的新闻语料共计 3581 篇。

对本过滤系统的性能我们采用 4 种评价方法做评估：前两种评价是在检索与过滤中应用最为广泛的准确率与召回率，另两种评价为 TREC2002 中定义的 T11F 和 T11U[6]。由于第二部分语料的相关文本数未知，因此，无法返回 T11F 与召回率测评结果。

5.2 结果分析与优化

表 1 为本系统的评测结果：

表 1 评测结果
Tab.1 evaluation result

	准确率	召回率	T11U	T11F
第一部分	0.667	0.778	21	0.686
第二部分	0.276	/	-5	/

在第一部分测试语料中，共过滤出文档 21 篇，其中相关文档 14 篇，关于重建却不相关的文档 5 篇。由结果可以看出系统对关于伊拉克重建方面文档准确率还是很高的，但是对更为具体的主题描述来说，系统对具体需求的理解上产生了偏差。

对未过滤出的 4 篇相关文档来说，导致内积值过低的因素主要有两个：第一个因素是文档的长度偏短，在计算特征项权重时，我们综合考虑了特征项的词频与段落频率，虽然对文档的长度作了归一化处理，但长度短的文档其段落层次也相对简单，总的来说，长文档比短文档中的特征项权重要大一些，致使那些较为简短的相关文档得出的内积值偏小；第二个因素是与特征项相关联的实例过少，对拥有“限定”属性的实例来说，只有在其“限定”属性值也匹配的情况下，特征项才与该实例相关联。但由于汉语中语法、句法格式的多变化，以及在构造本体时相关数据的稀疏问题，导致了即使特征项与实例的语义相同，也无法相关联的情况出现。在公式（2）中表现为 $S_i=0$ ，那么无论该项的权重有多大，也都不会对内积的值产生贡献，使得文档的内积值偏小。

综合这两方面因素，在公式（2）中增加特征向量的长度影响，并将特征项与实例相关联的要求降低。无论实例是否拥有“限定”属性，都与相应的特征项相关联，如果“限定”属性值也匹配的话，则增加该特征词的权重，使得 $d_i \cdot S_i$ 项的值增大。则改进的内积计算公式如下，其中， n 为特征项个数， k 为增加权重的特征项个数：

$$\text{RelationStrength} = \frac{k}{n} \sum_{i=1}^n d_i \cdot S_i \quad (3)$$

在第二部分的测试语料中，系统共过滤出 29 篇文档，其中完全符合用户主题需求的有 8 篇，涉及战后伊拉克问题或关于伊拉克重建的新闻共有 12 篇。对过滤出的不相关文档来说，既然拥有高于系统阈值的内积值，就说明该文档的所表述的语义与模板所表述的语义的相关程度符合系统的要求，但与用户的需求产生了偏差。从不相关文档的内容来看，导致问题产生的原因就在于文档的特征项的关联实例所属的类别分布不均匀。如果特征项关联的实例仅属于一个或少数几个类别的话，那么意味着本体模板中的某个类将拥有大量实例，而其他的类中实例仅有几个，甚至为 0。这样的实例分布虽然可能产生很高的关联值，但是此时体现的语义关联需求已经出现了偏差，不再符合用户的需求。另一个可能导致内积偏高的原因就是某一特征项的权重很大，使得的权重对内积的值的的影响增大，也可能导致需求的偏移，例如，在关于伊拉克的新闻中，“伊拉克”的权重通常比其他的特征词的权重大很多，如果权重过大的话，将影响到内积值所要表述的语义。为此，我们根据本体中的类对用户需求的不同贡献，分别赋予不同的权重，减小由于实例类别的分布不均衡或特征项权重过大而导致需求偏移的情况发生。在伊拉克重建中，用户最为关注的内容是伊拉克所面临的困难，以及提出的各种解决方案，因此，“困难”类和“方案”类中所受关注的程度应大于其它类。在公式（3）的基础上，改进内积计算公式：

$$\text{RelationStrength} = \sum_{j=1}^m \eta_j \cdot \left(\frac{k}{n+1} \sum_{i=1}^n d_{ij} \cdot S_{ij} \right) \quad (4)$$

其中 m 为本体模板中类的个数； n 为属于类 j 的实例数，将 n 加 1 是为了避免 n 取 0 的情况发生； η_j 为模板中的类别权重， $\sum_{j=1}^m \eta_j = 1$ 。

表 2 改进后测评结果

Tab.2 evaluating result after optimizing

	准确率	召回率	T11U	T11F
第一部分	0.80	0.889	28	0.816
第二部分	0.524	/	11	/

从表 2 可以看出, 将本体中的类赋予不同的权重可以很好的改进系统的过滤性能, 也可以将本体中的各个实例赋予不同的权值, 进一步细化用户的具体需求。

6. 结论与未来工作

由本体构造的用户需求模型, 利用本体中各节点间的关联保留了主题描述中的语义信息, 使得模板对主题需求的表述更为准确。实验结果显示, 本文的方法取得了较好的结果。本文的本体模板是根据某一具体需求构造的任务本体, 也可以将模板定义为一个领域的本体, 在同一领域本体内, 用户可以将本体划分为不同的区域来描述各自的不同需求, 这就使得用户可以在某一领域内对需求的维护变得格外的便利。但由于构造一个领域本体需要耗费相当的时间与资源, 因此在本文中并未涉及。另外, 对本体模板的动态更新与维护在本文也并未涉及, 这是我们下一阶段的工作内容。

参考文献:

- [1] 陈晓明, 王洪, 张仰森. “知网”的知识扩展和推理研究[J]. 贵州大学学报, 2001 年 5 月, 第 18 卷 第 2 期: 第 97 页.
- [2] Mustapha Baziz. Towards a Semantic Representation of Documents by Ontology-Document Mapping[J]. C. Bussler and D. Fensel (Eds.): AIMS 2004, LNAI 3192, pp. 33–43, 2004.
- [3] 林鸿飞, 麻志毅, 姚天顺. 基于语义框架的中文文本过滤模型[J]. 计算机研究与发展, 2001, Vol.38, 增刊, 136-141.
- [4] 晋耀红. 基于语义的文本过滤系统的设计与实现[J]. 计算机工程与应用, 2003 年 7 月, 第 22 页
- [5] Boanerges Aleman-Meza, Chris Halaschek, I. Budak Arpinar, and Amit Sheth ,Context-Aware Semantic Association Ranking[J], Technical Report 03-010, LSDIS Lab, Computer Science, University of Georgia, August 21, 2003
- [6] Stephen Robertson, Ian Soboroff. The TREC 2002 Filtering Track Report[J]. TREC2002 (SIGIR'02)
- [7] 王洪伟, 吴家春, 蒋 馥. 基于描述逻辑的本体模型研究[J]. 系统工程, 2003 年 5 月, 第 21 卷 第 3 期 第 101 页