

文章编号:1001-9081(2011)07-1751-05

doi:10.3724/SP.J.1087.2011.01751

带约束本体语义扩展的信息过滤算法

颜晶晶^{1,2}

(1. 台州职业技术学院 计算机工程系,浙江 台州 318000; 2. 同济大学 电子与信息工程学院,上海 201804)

(jingjing_yan@126.com)

摘要:提出一种基于本体的信息过滤方法。该方法通过本体实现形式化语义描述,并对原始输入条件进行带约束规则的本体语义扩展。进而为了实现语义匹配,给出了信息向量语义描述及权重计算方法。最终,实现基于语义相似度计算的信息过滤。实验证明,该方法是有效的。

关键词:本体;语义扩展;约束;相似度;信息过滤

中图分类号:TP391 **文献标志码:**A

Mechanism of ontology semantic extension with constraints for information filtering

YAN Jing-jing^{1,2}

(1. Department of Computer Engineering, Taizhou Vocational and Technical College, Taizhou Zhejiang 318000, China;

2. School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract: An ontology based information filtering method was proposed in this paper. This approach described formal semantics through ontology so that user's original input query terms could be converted into ontology semantic extension formats with constraint rules. Furthermore, semantic vector definition of information and its right computation method were presented in order to achieve the semantic matching between user's requirements and information. Finally, the semantic similarity computation based information filtering was addressed. The experimental results show that the proposed information filtering algorithm is effective.

Key words: ontology; semantic extension; constraint; similarity; information filtering

0 引言

面对互联网上信息和数据的海洋,人们往往因为无法快速准确获取信息而束手无策,“信息过载”、“信息爆炸”问题随处可见。以搜索引擎为代表的信息检索技术为人们提供了一种信息获取方案^[1-2]。以关键词检索为核心的搜索引擎进行准确信息获取的前提是:用户可以准确地将自己内心信息获取意图转化为关键词输入。然而,这一前提往往无法获得满足。当用户输入较为模糊时,信息获取结果将会庞杂冗余,人们不得不对返回结果进行二次检索。如何准确有效地为系统提供用户需求,实现有深层次且针对性的信息服务,是当前信息检索领域一个重要的课题。

我们认为,信息获取的关键在于将用户需求的语义与信息所包含的语义进行准确比较,使系统在理解这两者语义的基础上进行信息过滤,去除无关信息,就可以为用户提供真正准确的信息服务。因此,将信息获取和过滤建立在计算机自动处理语义的基础上,是一种可行且有效的方法。本体技术正可以为我们提供这样的解决方案^[3]。

信息过滤技术历经国内外许多学者的研究,已经取得了众多成功。AT&T实验室的 William W. Cohen 提出了一种邮件分类规则学习方法,利用基于 RIPPER 规则学习算法和关键词学习规则来进行邮件分类^[4]。Salvador Nietosanchez 等人将 OCAT(One Clause at A Time)挖掘算法用于文本分类^[5-6],根据正反例集合训练得到规则集合,并根据规则判断新文本。

文献[7]根据用户提供的术语和反馈,建立中介索引,分析时间和经验因素,并对从搜索引擎得到的结果进行过滤,从而发现用户真正需要的信息。与此同时,国内众多学者也提出了多种信息过滤方案^[8-11]。然而这些方案所采用的方法大多数存在一定的局限:依赖于大量语料库的辅助,对前期信息过滤的样本训练依赖度大,过滤模型的知识依赖于人工建立等。

针对上述问题,本文提出了一种基于本体语义扩展的信息过滤算法,利用领域本体进行语义描述,从而使计算机具备“读懂”和自动处理语义的能力。进而,提出利用本体的良好知识结构进行推理,实现语义扩展,为用户输入的检索条件添加“潜在语义”。然后,给出了信息语义的形式化描述方式,并在此基础上实现信息语义过滤。

1 本体及语义扩展

本文定义出一种领域本体用以给出语义描述手段以及可推理的知识结果。

定义 1 信息过滤中的本体可以表示为四元组 $O = (C, R, P, F)$ 。其中: C 表示概念名集合, R 表示概念之间的关系集合, P 表示描述概念的属性集合, F 表示本体中概念、属性等元素之间存在的约束关系集合。

根据上述定义,我们可以从本体中获得如下语义描述:

1) 关系语义。若概念 c_1 和概念 c_2 之间存在关系 r_1 , 则可获得语义描述 $r_1(c_1, c_2)$ 。例如概念“生物”与“动物”之间为“父子关系”, 则本体所表达的语义为 $parent(\text{生物}, \text{动物})$ 。

2) 描述语义。若概念 c_1 存在属性集合 P^1 , 则具有语义描述 $c_1.p_i^1$ 。

3) 约束语义。若存在约束式 $f_j \in F$, 则所有约束式 f_j 所涉及的本体元素 $x \in O$ 均需服从 f_j , 记为 $f_j \rightarrow x$ 。例如存在约束式 $f(c_1) \rightarrow c_2$, 则表示存在关于概念 c_1 的约束规则 f 表明概念 c_2 在推理时与 c_1 互斥。约束语义在后面进行语义扩展时很重要。

在本体语义描述的基础上, 给出语义扩展的具体定义。

定义 2 本体语义扩展是指如下情况之一:

1) 概念语义扩展: 对于一个概念 c' 来说, 若存在概念 $c'' \in O.C$, 具有概念关系语义且满足 $c'' = \{x | R(c', x) \vee R(x, c')\}$, 则称概念 c'' 为概念 c' 的概念语义扩展, 概念语义扩展得到的集合记为 $R(c')$ 。

2) 描述语义扩展: 对于概念 c' 具有已给定的描述属性集合 S , 若在本体中存在描述语义 $p_i \in O.c'.P \wedge p_i \notin S$, 则称属性 p_i 为概念 c' 的描述语义扩展, 描述语义扩展得到的集合记为 $R(c'.P)$ 。

然而这种语义扩展是一种基本的遍历方法, 仅仅有助于计算机实现类似于“概念联想”之类的检索, 可以实现联想式检索, 在查全率、关联度提升上有一定的帮助。实际上, 这反而会导致信息冗余。因此, 我们结合本体约束, 提出带约束的本体语义扩展。

定义 3 带约束的本体语义扩展是指如下情况之一:

1) 约束概念语义扩展: 对于一个概念 c' 及其概念语义扩展集合 $R(c')$ 而言, 若存在与其相关的规则集合 $L(c') \subseteq O.F$, 则在该规则集合中的任意一条规则 $\xi_k \in L(c')$ 约束下, 其扩展概念应满足 $c'' \in R(c') \wedge \xi_k(c') \rightarrow c''$, 记为 $L;R(c')$ 。

2) 约束描述语义扩展: 对于一个概念 c' 及其描述语义扩展集合 $R(c'.P)$ 而言, 若存在与其相关的规则集合 $L(c'.P) \subseteq O.F$, 则在该规则集合中的任意一条规则 $\xi_k \in L(c'.P)$ 约束下, 其扩展概念应满足 $p_i \in R(c'.P) \wedge \xi_k(c'.P) \rightarrow p_i$, 记为 $L;R(c'.P)$ 。

图 1 中给出了各类概念语义扩展和描述语义扩展的示例。

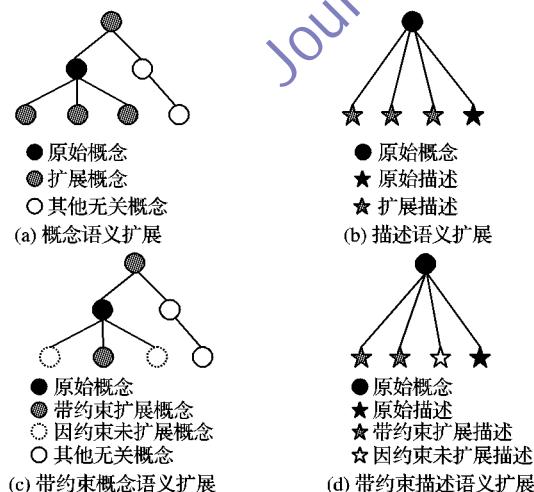


图 1 本体语义扩展示例

下面我们给出本体语义扩展的算法。

输入: 概念集合 D , 以及概念 $d_i \in D$, 属性集合 $d_i.Q$;
输出: 约束语义扩展集合 $\xi_k:R(D), \xi_k:R(D.Q)$ 。

- 1) $\emptyset \rightarrow Y; \emptyset \rightarrow Z$;
- 2) For each $d_i \in D$ do

- 3) If $\text{find}(c_j \in O \wedge (\exists r(d_i, c_j) \vee \exists r(c_j, d_i)))$ then $c_j \in R(d_i)$;
- 4) If $\text{find}(\xi_k \in O.F \wedge (\exists \xi_k \rightarrow d_i))$ then $\xi_k:R(d_i) \rightarrow Y$;
- 5) For each $d_i \in Q$ do
- 6) If $\text{find}(\exists (q_l \in O.d_i.P) \wedge (q_l \notin d_i.Q))$ then $q_l \in R(d_i.Q)$;
- 7) If $\text{find}(\xi_k \in O.F \wedge (\exists \xi_k \rightarrow d_i.Q))$ then $\xi_k:R(d_i.Q) \rightarrow Z$;
- 8) Return Y, Z

2 用户查询条件权重计算

我们对用户输入的查询条件语义进行定义, 分为原始条件语义和扩展条件语义两部分。原始条件语义是用户直接输入的部分, 体现用户最为关心的核心问题; 扩展条件语义由上述语义扩展方法得到, 体现用户可能感兴趣的语义部分。通过权重计算方法为系统标注出用户关心的重点, 从而尽量使信息获得结果与用户意图关联度高。

定义 4 用户查询条件语义可以表示为二元组: $K = (M, E)$, 其中 $M = (C_M, P_M)$ 为原始查询条件, C_M 为原始查询概念集合, P_M 为原始查询属性集合; $E = (L; R(C_M), L; R(C_M, P_M))$ 为扩展查询条件, $L; R(C_M)$ 为约束概念语义扩展集合, $L; R(C_M, P_M)$ 为约束描述语义扩展集合。

定义 5 本体权重。对于本体中的概念和属性, 其权重 rc 和 rp 分别如下描述:

- 1) 本体根概念节点的权重为 1;
- 2) 若存在概念 c' 与概念 c'', c'' 是 c' 的直接子概念, 则满足 $rc' = \sum rc''$;
- 3) 概念 c_i 存在 m 个描述属性 $c'.p_j$, 则对于一个概念的所有属性而言, 其权重满足 $\sum_{j=1}^m rp_j = 1$ 。

根据定义 4 和定义 5, 可以得到用户查询条件语义中各元素的权重。为了区别原始查询和扩展查询的重要性, 对这两类查询条件权重分别计算。

用户训练样本集 X 记录了用户过往所有查询条件输入, 假设存在过往原始输入中概念 c_{Mi} 出现过 $g(c_{Mi})$ 次, 属性 p_{Mj} 出现过 $g(p_{Mj})$ 次; 而样本集中共出现原始输入概念 n 次, 原始输入属性 m 次, 则原始查询条件的概念 c_{Mi} 和属性 p_{Mj} 分别可获得权重为:

$$\begin{cases} \omega_{c_{Mi}} = rc_{Mi} \times \beta^{\frac{g(c_{Mi})}{n}} \\ \omega_{p_{Mj}} = rp_{Mj} \times \beta^{\frac{g(p_{Mj})}{m}} \end{cases} \quad (1)$$

其中 β 为调节参数, 且 $0 \leq \beta \leq 1$, 其作用为使概念或属性在训练样本集中出现次数越多可获得的重要度越高。

与原始查询条件类似, 我们给出了扩展查询条件中概念和属性权重计算方法。对于扩展概念 $c_{Ei} \in L; R(C_M)$ 以及扩展属性 $p_{Ej} \in L; R(C_M, P_M)$ 而言, 假设存在过往扩展概念 c_{Ei} 出现过 $g(c_{Ei})$ 次, 属性 p_{Ej} 出现过 $g(p_{Ej})$ 次; 而样本集 X 中共出现扩展概念 u 次, 扩展属性 v 次, 则扩展查询条件的概念 c_{Ei} 和属性 p_{Ej} 分别可获得权重为:

$$\begin{cases} \omega_{c_{Ei}} = rc_{Ei} \times \beta^{\frac{g(c_{Ei})}{u}} \\ \omega_{p_{Ej}} = rp_{Ej} \times \beta^{\frac{g(p_{Ej})}{v}} \end{cases} \quad (2)$$

3 信息的概念语义权重计算

信息通过自然语言描述, 但是其关键内容仍然可以通过

概念及其属性表达出来。例如许多文章会给出标签,则这些标签就是信息的关键内容。利用基于向量的概念和属性方法来表示信息的主要内容。

定义6 向量语义。向量语义可表示为: $\rho = (c, (c.p_1, c.p_2, \dots, c.p_t), right)$, 其含义为由概念 c 及其描述属性集合 $(c.p_1, c.p_2, \dots, c.p_t)$ 以及其重要程度 $right$ 所组成的语义集合。

定义7 信息向量语义。一条信息语义可以表示为向量 $\Pi = (\Phi, \Omega)$, $\Phi = (\rho_1^\phi, \rho_2^\phi, \dots, \rho_r^\phi)$ 中向量语义 $\rho.right \geq \mu_1$ (μ_1 为预设阈值), 表示 ρ 为信息中关键内容所组成的主语义; 而 $\Omega = (\rho_1^\omega, \rho_2^\omega, \dots, \rho_s^\omega)$ 中向量语义 $\rho.right < \mu_1$, 表示 ρ 为信息较为次要的内容所组成的副语义。

下面给出向量重要度 $right$ 的计算方法。假设一条信息经分解后,其中共有 h_c 个概念且这些概念共出现 d_c 次,共有 h_p 个属性且这些属性共出现 d_p 次(概念及属性可以在不同位置多次出现,这里的次数为所有位置上出现次数的总和)。借鉴文献[12]的思想并对其计算方法进行改进,将向量重要程度分为两个方面计算。

1) 频率重要度 $right_{freq}$ 。该指标表达了向量语义在信息中出现次数所体现出的重要性。

$$right_{freq}(\rho_i) = \alpha \times \left(\rho_i \cdot rc \times \frac{|\rho_i \cdot c|}{d_c} \right) + (1 - \alpha) \times \sum_{j=1}^t \left(\rho_i \cdot c \cdot rp_j \times \frac{|\rho_i \cdot cp_j|}{d_p} \right) \quad (3)$$

其中: $|\rho_i \cdot c|$ 表示向量语义 ρ_i 的概念 c 在信息中出现的次数,参数 $\alpha (0 < \alpha < 1)$ 为调节参数。

2) 位置重要度 $right_{loc}$ 。该指标表达了向量语义在信息中出现的位置所占的重要性。假设文档中存在 l 个不同指定位置 χ , 每一个位置本身的权重为 $\delta_k^{x_k}$, 且满足 $\sum_{k=1}^l \delta_k^{x_k} = 1$ 。向量语义 ρ_i 在第 χ_k 个位置出现的次数表示为 $|\rho_i^{x_k}|$, 而在该位置上共出现概念 $d_c^{x_k}$ 次和属性 $d_p^{x_k}$ 次。向量语义 ρ_i 在信息中所占的位置重要度 $right_{loc}$ 可计算为:

$$right_{loc} = \alpha \times \left[\sum_{k=1}^l \left(\delta_k^{x_k} \times \frac{|\rho_i \cdot c^{x_k}|}{d_c^{x_k}} \right) \right] + (1 - \alpha) \times \left[\frac{\sum_{k=1}^l \delta_k^{x_k} \times \left(\sum_{j=1}^t \frac{|\rho_i \cdot cp_j^{x_k}|}{d_p^{x_k}} \right)}{l \times t} \right] \quad (4)$$

依据上述两类重要度计算,可以获得向量语义的 ρ_i 的重要度 $right$:

$$right = \frac{right_{freq} + right_{loc}}{2} \quad (5)$$

4 带约束本体语义扩展的信息过滤算法

4.1 信息语义与用户查询条件相似度计算

信息过滤的关键就在于根据用户输入条件从候选信息集合中挑选出最符合用户意图的信息反馈给用户。因此,在这个过程中必须在信息语义用户查询条件语义之间建立有效的相似度计算方法。

1) 概念相似度。设本体中存在概念 c_1 和 c_2 , 则这两个概念的相似度为:

$$\text{sim}(c_1, c_2) =$$

$$\begin{cases} 1, & c_1 \text{ 和 } c_2 \text{ 为同一概念} \\ \frac{1}{\text{path}(c_1, c_2)}, & c_1 \text{ 和 } c_2 \text{ 为祖先—后代关系} \\ 0, & \text{其他关系} \end{cases} \quad (6)$$

式(6)中, $\text{path}(c_1, c_2)$ 表示概念之间的路径中含有的节点总数。例如: 概念 c_1 是 c_2 的父亲节点, 则 $\text{path}(c_1, c_2) = 2$; 若存在 c_3 为 c_2 的儿子节点, 则 $\text{path}(c_1, c_3) = 3$ 。但该函数仅计算祖先—后代关系。

2) 属性相似度。设本体中存在两个属性 p_1 和 p_2 , 则这两个属性之间的相似度为

$$\text{sim}(p_1, p_2) =$$

$$\begin{cases} 1, & p_1 \text{ 和 } p_2 \text{ 属于同一概念且 } p_1 = p_2 \wedge F(p_1) : p_2 \\ 0.5, & p_1 \text{ 和 } p_2 \text{ 属于同一概念且 } p_1 = p_2 \wedge \exists (\neg F(p_1)) : p_2 \\ 0, & \text{其他关系} \end{cases} \quad (7)$$

上述式的语义为若属性 p_1 和 p_2 属于同一个概念且这两个属性相同,同时 p_2 属性值符合 p_1 的约束条件,则这两个属性相似度为 1; 若属性 p_1 和 p_2 属于同一个概念且这两个属性相同,同时存在 p_2 属性值不符合 p_1 的约束条件之一,则这两个属性相似度为 0.5, 其他情况为 0。

3) 原始查询条件。原始查询条件体现了用户的核心意图,因此我们将原始查询条件与信息语义的主语义及副语义同时比较。设原始查询条件 $M = (C_M, P_M)$, 信息语义为 $\Pi = (\Phi, \Omega)$, 则两者的相似度可计算为:

$$\text{sim}(M, \Pi) = \gamma \times \left[\frac{\sum_{i=1}^{|C_M|} \max(\text{sim}(c_M^i, \rho_i^\Pi))}{|C_M|} \right] + (1 - \gamma) \times \left[\frac{\sum_{j=1}^{|P_M|} \max(\text{sim}(c_M \cdot p_j, \rho_i^\Pi \cdot p_j))}{|P_M|} \right] \quad (8)$$

式中,参数 $\gamma (0 < \gamma < 1)$ 为调节参数。 $\max(\text{sim}(c_M^i, \rho_i^\Pi))$ 的含义可以表示为:

$$\max(\text{sim}(c_M^i, \rho_i^\Pi)) = \max[\text{sim}(c_M^i, \rho_1^\phi), \dots, \text{sim}(c_M^i, \rho_r^\phi), \text{sim}(c_M^i, \rho_1^\omega), \dots, \text{sim}(c_M^i, \rho_s^\omega)] \quad (9)$$

同理, $\max(\text{sim}(c_M \cdot p_j, \rho_i^\Pi \cdot p_j))$ 为类似含义。

4) 扩展查询条件。扩展查询条件为用户核心意图的一种衍生,因此扩展查询条件仅与信息语义的主语义比较。设扩展查询条件 $E = (L:R(C_M), L:R(C_M, P_M))$, 信息主语义为 $\Phi = (\rho_1^\phi, \rho_2^\phi, \dots, \rho_r^\phi)$, 则两者的相似度可计算为:

$$\text{sim}(E, \Phi) = \gamma \times \left[\frac{\sum_{i=1}^{|C_E|} \max(\text{sim}(c_E^i, \rho_i^\Phi))}{|C_E|} \right] + (1 - \gamma) \times \left[\frac{\sum_{j=1}^{|P_E|} \max(\text{sim}(c_E \cdot p_j, \rho_i^\Phi \cdot p_j))}{|P_E|} \right] \quad (10)$$

4.2 带约束本体语义扩展的信息过滤算法

假设候选信息集合含有若干个信息语义 $\Pi_y (y = 1, 2, \dots, n)$, 用户原始查询输入为 $K = M$ 。本文采用如下信息过滤流程:

1) 对原始查询输入 $K = M$ 进行语义扩展,获得 $K = (M, E)$;

2) 计算 $\omega_{C_M}, \omega_{P_M}$, 计算 $\omega_{C_E}, \omega_{P_E}$;

- 3) 对于每一个 Π_y , 计算 $\Pi_y \cdot \rho_i \cdot right$;
 4) 对于每一个 Π_y , 计算 $sim(M_i, \Pi_y)$ 和 $sim(E, \Pi_y, \Phi)$;
 5) 计算:

$$\varpi_1 = \frac{\sum \omega_{cMi} + \sum \omega_{pMi}}{3} + sim(M, \Pi_y) \quad (11)$$

6) 计算:

$$\varpi_2 = \frac{\sum \omega_{cEi} + \sum \omega_{pEi}}{3} + sim(E, \Pi_y, \Phi) \quad (12)$$

7) 计算用户查询条件与信息语义之间的相似度:

$$sim(K, \Pi_y) = \sqrt{\varpi_1^2 + \varpi_2^2} \quad (13)$$

8) 依据 $sim(K, \Pi_y)$ 形成信息关联度队列, 对于预设阈值 μ_2 , 删除队列中的 Π_y 满足 $sim(K, \Pi_y) < \mu_2$ 。

5 实验分析

为了验证本文的工作, 自主开发了医药领域文本案例过滤系统原型以及医药领域案例本体。该系统包括自主开发的中文词语分析系统以及案例过滤系统两个部分。医药领域案例本体采用 Protégé 4 工具开发; 中文分词系统在 Eclipse 下采用 Java 开发; 案例过滤系统原型采用 Java 开发。实验中有关公式计算的取值如下: 所有预设参数及阈值取值为 0.6; 公式中的参数 $\beta = 0.9, \alpha = 0.6, \gamma = 0.6$ 。候选案例集来源于网络中所搜集并预处理的 200 个医药行业案例, 分为医药制造、医药疗效、医药销售、医药广告等 4 个大类。同时, 加入与上述 4 个大类无任何关联(但带有医药领域概念或属性词汇出现)的噪声案例 20 个。

5.1 带约束的语义扩展准确性测试

准确性测试是为了验证本文方法是否可以准确地实现信息过滤。分为两组实验, 每组进行 10 次信息过滤, 每次从 200 个案例中随机选取 150 个信息作为候选过滤集。Group1 中候选集不带有噪声信息。Group1 分为两组, Baseline1 采用原始关键词(即关键词匹配技术, 与本文所采用的原始查询条件方法不同。区别在于: 关键词匹配方法直接将输入条件与文献中的词汇进行比较匹配, 而本文所述原始查询条件已经将原始查询条件进行语义化处理, 形成 $M = (C_M, P_M)$ 的形式化语义)进行过滤, Line1 采用本文方法进行过滤。准确性评价指标计算如下:

$$ratio1 = \frac{\text{过滤后用户可接受信息数量}}{\text{用户可接受信息数量}} \quad (14)$$

其中: 用户可接受信息数量为我们在每次过滤前人工选择出来形成的集合, 过滤后用户可接受信息数量为人工统计方法获得。从图 2 中可以看到, 采用本文过滤方法以后, 用户可接受信息比例明显高于参照组 Baseline1。这是由于 Line1 采用了语义扩展方法, 扩展出来的语义增加了用户可接受的信息数量, 因此 $ratio1$ 明显高于 Baseline1。

Group2 也进行 10 次过滤, 除了每次从 200 个案例中随机选取 150 个信息作为候选过滤集之外, 将 20 个噪声信息加入候选集中。Baseline2 采用原始关键词(与上一实验方法一样)过滤, Line2 采用本文方法进行过滤。除了继续采用 $ratio1$ 作为评价指标外, 还引入噪声过滤准确率指标, 计算如下:

$$ratio2 = \frac{\text{被删除的噪声信息}}{\text{噪声信息总数}} \quad (15)$$

由图 3 中可以看出 Group2 中参照组 Baseline2 没有语义约束, 原始查询条件语义清晰度不高; 而 Line2 中由于采用了语义约束方法, 准确率明显提高; 同时在图 4 中可以看出, 对于噪声信息, 约束规则的过滤效率明显提升。

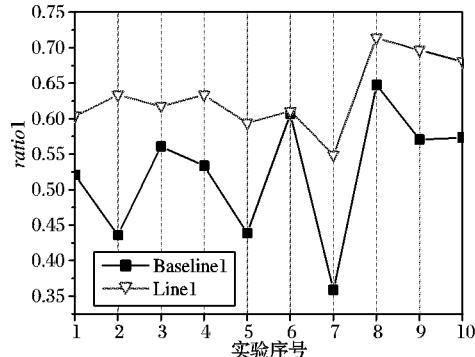


图 2 不带噪声候选案例集的过滤准确性测试 (Group1)

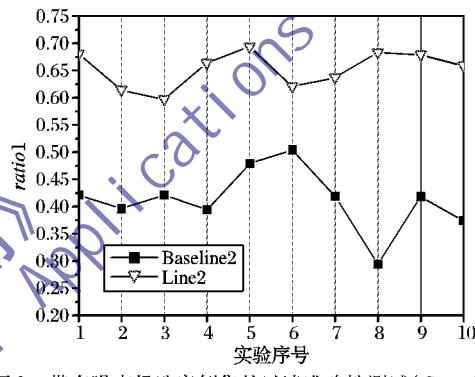


图 3 带有噪声候选案例集的过滤准确性测试 (Group2)

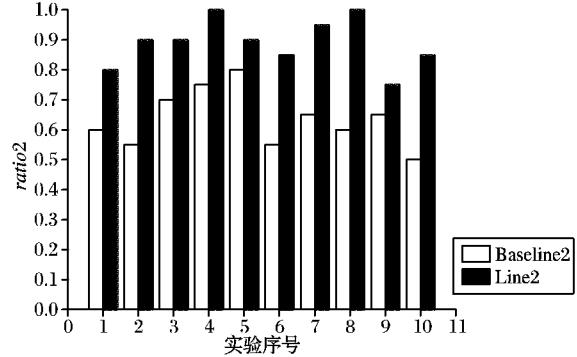


图 4 带有噪声的候选案例集噪声过滤准确性测试 (Group2)

5.2 带约束的语义扩展效果测试

约束扩展的作用在于可以更好地为过滤提供规则。因此, 设计了 Group3 的对比实验。测试采用三种方法进行: Baseline3 为无扩展方法的过滤, 直接采用原始查询条件进行过滤; Baseline4 为无约束语义扩展过滤, 仅对原始条件进行本体语义扩展, 不采用约束规则; Line3 为本文所提出的带约束语义扩展效果。每种方法测试均进行 20 次过滤, 候选案例集为带有噪声案例的 220 个案例库。记录了测试后式 (14) 计算所得指标值累加后的算术平均值, 以跟踪每种方法的平均效果。

图 5 给出了三种方法实验的结果, 可以看出本文方法明显优于另外两种方法, 可见采用语义扩展后可能显著改善过滤后用户感兴趣的信息获取效果, 这主要是因为扩展后可将大量原始输入条件语义中没有表达出的潜在需要扩展出来; 而通过约束条件, 又可以进一步提升扩展精度, 从而降低无约束扩展带来的冗余可能性。

6 结语

信息过滤是人们从海量信息中获取真正所需信息的重要检索技术。但是,信息过滤长期以来受困于精度不高、无法真正实现信息内容语义检索等困难。本文引入本体技术,利用本体对信息进行形式化语义描述,从而使信息与用户要求之间存在可计算的语义基础。同时,我们利用本体实现带约束的语义扩展,将用户未能明确输入的潜在语义扩展出来,并用约束规则对其进行规定,从而避免带来巨大冗余。最终,本文给出了带约束本体语义扩展条件下的信息过滤算法,实现相似度计算下的信息过滤。在以后的工作中,将重点针对形式化约束下的推理展开工作,从而进一步提高语义扩展的精度。

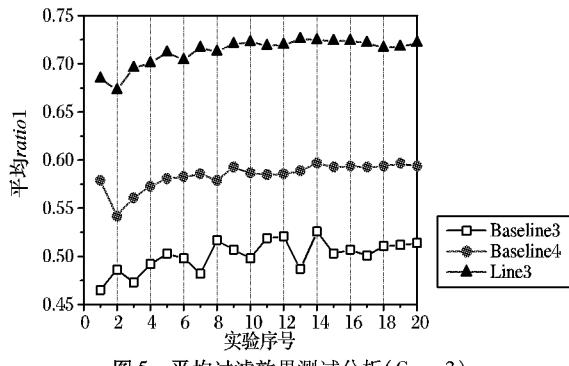


图5 平均过滤效果测试分析(Group3)

参考文献:

- [1] BELKIN J, BRUCE CROFT W. Information filtering and information retrieval: Two sides of the same coin [J]. Communications of the ACM, 1992, 35(12): 29 – 38.
- [2] BHANDARKAR S M, LUO XING-ZHI. Integrated detection and tracking of multiple faces using particle filtering and optical flow-based elastic matching[J]. Computer Vision and Image Understanding, 2009, 113(6): 708 – 725.
- [3] GRUBER T R. A translation approach to portable ontology specification [J]. Knowledge Acquisition, 1993, 5(2) : 199 – 220.
- [4] LI KANG, ZHONG ZHEN-YU, LAKSHMISH R. Privacy-aware collaborative spam filtering [J], IEEE Transactions on Parallel and Distributed Systems, 2009, 20(5): 725 – 739.
- [5] DESHPANDE A S, TRIANTAPHYLLOU E. A greedy randomized adaptive search procedure (CRASP) for inference logical clauses from examples in polynomial time and some extensions[J]. Mathematical and Computer Modelling, 1998, 27(1): 75 – 99.
- [6] SANCHEZ S N, TRIANTAPHYLLOU E, KRAFT D H. A feature mining based approach for the classification of text documents into disjoint classed information[J]. Information Processing and Management, 2002, 38(4): 583 – 604.
- [7] DREILINGER D, HOWE A E. Experiences with selecting search engine using metasearch[J]. ACM Transactions on Information Systems, 1997, 15(3) : 195 – 222.
- [8] 曾春, 邢春晓, 周立柱. 基于内容过滤的个性化搜索算法[J]. 软件学报, 2003, 14(5): 999 – 1004.
- [9] 田范江, 李从蓉, 王鼎兴. 进化式信息过滤方法研究[J]. 小型微型计算机系统, 2003, 11(3): 328 – 333.
- [10] 梁理, 黄樟钦, 侯义斌. 网络信息过滤系统(NFIS)的研究与实现[J]. 小型微型计算机系统, 2003, 24(2): 195 – 198.
- [11] 张波, 向阳, 王坚. 一种基于语义可理解的信息过滤算法[J]. 电子与信息学报, 2010, 32(10): 2324 – 2330.

(上接第1750页)

统算法,这是由于我们在预测评分时增加了对邻居用户信任度的度量,体现了不同邻居用户评分对最终目标用户推荐的贡献程度,尤其是引入的属性偏好度函数,使得传统算法中无法解决的新项目得到了有效推荐,从而进一步提高了预测的可靠性和系统的推荐性能。

4 结语

针对传统协同过滤算法中评分数据的极度稀疏及冷启动问题,本文提出了基于用户特征和项目属性的协同过滤推荐算法,通过衡量用户的兴趣变化、受信任程度及对项目不同属性的偏好程度来预测目标用户的评分。相比仅仅依赖用户评分数据寻找最近邻居并进行推荐的传统算法,本文算法能有效改善邻居识别的准确性和预测的可靠性,提高整个系统的推荐质量。

参考文献:

- [1] SCHAFER J B, KONSTAN J A, RIEDL J. E-commerce recommendation applications[J]. Data Mining and Knowledge Discovery, 2001, 5 (1/2): 115 – 153.
- [2] SARWAR B, KARYPIS G, KONSTAN J, et al. Analysis of recommendation algorithms for e-commerce[C]// ACM Conference on Electronic Commerce. New York: ACM, 2000: 158 – 167.
- [3] HERLOCKER L J, KONSTAN A J, RIEDL T J. Empirical analysis of design choices in neighborhood-based collaborative filtering algorithms[J]. Information Retrieval, 2002, 5(4) : 287 – 310.
- [4] BREESE J, HECHERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering[EB/OL]. [2010 – 10 – 05]. <http://www.cs.pitt.edu/~mrotnar/comp/rs/Breese%20UAI%201998.pdf>.
- [5] 孙小华. 协同过滤系统的稀疏性与冷启动问题研究[D]. 杭州: 浙江大学, 2005.
- [6] HERLOCKER J L, KONSTAN J A, BORCHERS A, et al. An algorithmic framework for performing collaborative filtering[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1999: 230 – 237.
- [7] WANG JUN, de VRIES A P, REINDERS M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion[C] // Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2006: 501 – 508.
- [8] 王茜, 王均波. 一种改进的协同过滤推荐算法[J]. 计算机科学, 2010, 37(6): 226 – 228.
- [9] 张富国. 用户多兴趣下基于信任的协同过滤算法研究[J]. 小型微型计算机系统, 2008, 29(8) : 1415 – 1419.
- [10] 李聪, 梁昌勇. 基于属性值偏好矩阵的协同过滤推荐算法[J]. 情报学报, 2008, 27(6): 884 – 890.
- [11] HERLOCKER L J, KONSTAN A J, TERVEEN G L, et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems, 2004, 22(1): 5 – 53.
- [12] KARYPIS G. Evaluation of item-based top-n recommendation algorithms[C]// Proceedings of the 10th International Conference on Information and Knowledge Management. New York: ACM, 2001: 247 – 254.