第 29 卷 第 1 期
2007 年 1 月

武 汉 理 工 大 学 学 报
JOURNAL OF WUHAN UNIVERSITY OF TECHNOLOGY

Vol. 29 No. 1
Jan. 2007

语义相似性与协同过滤集成推荐算法研究

罗耀明,聂规划
(武汉理工大学管理学院, 武汉 430070)

摘 要: 基于项目协同过滤算法能提高基于用户协同过滤方法的扩展性问题,并考虑项目之间的关系避免计算用户之间关系的瓶颈,但基于项目协同过滤算法依然存在稀疏性和新项目预测等问题。为了解决这些问题,该文采用了一种基于项目的结构化语义信息的集成相似性算法。为了抽取项目的语义信息,通过本体学习建立特定领域本体并利用包装器代理从网站中抽取本体类的实例和项目属性。实验结果证明了此方法不仅能很好的解决基于项目协同过滤算法带来的问题,而且还提高了推荐精度。

关键词: 推荐系统; 协同过滤; 语义相似性; 本体
中图分类号: TP 301.5 文献标志码: A 文章编号: 1671-4431(2007)01-0085-04

Research of Recommendation Algorithm on Integration of Semantic Similarity and the Item-based CF

LUO Yao-ming, NIE Gui-hua
(1. School of Management, Wuhan University of Technology, Wuhan 430070, China)

Abstract: Item-based Collaborative Filtering algorithms can enhance the scalability problems associated with traditional user-based Collaborative Filtering approaches and avoid the bottleneck of computing user-user correlations by considering the relationships among items. But it still worked poor in solving the problem of sparsity, predictions for new Items. In order to resolve efficiently several problems, this paper introduced an integrated similarity algorithms based on structured semantic knowledge about Items. We built domain-special ontology by ontology learning and used wrapper agents to automatically extracting instances of the ontology classes and semantic properties about Items from web site. Experimental results showed that the integrated similarity algorithms efficiently deal with the problems associated with Item-based Collaborative Filtering algorithms as well as improving accuracy.

Key words: recommendation systems; collaborative filtering; semantic similarity; ontology

随着 Internet 的发展和电子商务的应用,网上已出现信息过载的现象,于是产生了协同过滤技术的个性化推荐系统。传统推荐系统的基本思想是基于评分相似的最近用户邻居的评分数据向目标用户产生推荐[1,2]。尽管传统的推荐系统是当前使用最成功的技术,但它也有许多不足之处[3]。基于项目协同过滤算法[4]的提出虽然避免了传统的协同过滤算法计算用户之间相似性的瓶颈,但依然存在一些缺陷。关键性问题是:每个用户一般都只对很少的项目感兴趣,整个用户评分数据非常稀疏,这就导致用户之间的相似性不准确,产生的最邻近的邻居用户不可靠;难以推荐或预测一个新项目。该文提出了一种基于语义相似性的项目协同过滤算法,该方法将基于用户平分计算项目的相似性与语义相似性组合,能很好地解决上述问题。

收稿日期: 2006-10-15.
基金项目: 国家自然科学基金(70572079).
作者简介: 罗耀明(1979-), 男, 博士生. E-mail: luoyaoming@sohu.com

1 基于项目的协同过滤算法

基于项目的协同过滤推荐根据用户对相似项目的评分预测该用户对目标项目的评分，基于该假设：如果大部分用户对一些项目的评分比较相似，则当前用户对这些项目的评分也比较相似。基于项目的协同过滤推荐系统使用统计技术找到目标项目的若干最近邻居，由于当前用户对最近邻居的评分与对目标项目的评分比较类似，可以根据当前用户对最近邻居的评分预测当前用户对目标项目的评分，产生对应的推荐列表。

1)项目相似性计算　在基于项目协同过滤算法中，关键是计算项目之间的相似性，然后选择最相似的项目。现在有许多不同的计算项目之间的相似性的方法[3]，例如：余弦相似性、相关相似性、修正的余弦相似性，下面介绍修正的余弦相似性的方法。

修正的余弦相似性：设 $S(i,j)$ 表示资源 i 与资源 j 之间的相似性，项目 i 和项目 j 共同评过分的用户集合用 U 表示，则项目 i 和项目 j 之间的相似性 $S(i,j)$ 为

$$S(i,j)=\frac{\sum_{C\in U}(R_{C,i}-R_C)(R_{C,j}-R_C)}{\sqrt{\sum_{C\in U}(R_{C,i}-R_C)^2}\sqrt{\sum_{C\in U}(R_{C,j}-R_C)^2}} \tag{1}$$

$R_{C,i}$ 表示用户 C 对项目 i 的评分，R_C 表示用户 C 对项目的平均评分。

2)预测计算　在计算项目之间的相似性之后，要选择 k 个与目标项目最相似的项目，并产生目标项目的预测值。采用相似资源评价的权重组合方法，生成用户对目标资源的预测评价值。其计算式为

$$P_{a,i}=\sum_{j=1}^{k}(P_{a,j}\times S(i,j))/\sum_{j=1}^{k}S(i,j) \tag{2}$$

$P_{a,j}$ 表示用户 a 对目标项目 j 的预测值，这里仅有 k 个预测值产生。

3)基于项目的协同过滤算法不足　基于项目的协同过滤算法通过计算项目之间的相似性，选择与目标项目的最近邻居集合，避免了计算用户之间相似性的瓶颈，该算法比基于用户协同过滤算法的扩展性强，精确度高。但还是存在数据稀疏性和新项目预测的问题。为了解决这 2 个问题，将采用基于语义相似性过滤算法。在算法中，从网页中自动地抽取结构化的项目语义知识，并结合项目评分组合一种新的项目相似性方法。

2 基于本体的语义相似性的协同过滤算法

在协同过滤处理过程中，为了获取项目的语义信息，必须从多个网页或网站中抽取具有概念层次的结构化对象作为语义实体。但现在的网络，其信息存储为静态 HTML 页面，用于表达 Web 页面信息的 HTML 标记语言存在着缺点：HTML 语言的标记(tag)只是告诉浏览器如何显示它所定义的信息，却不包含语义信息。针对这个问题引入了本体技术。

1)本体与本体的学习　一个本体提供了定义好的重要概念以及概念之间的语义关系的结构，Ontology 的目标是捕获相关领域的知识，提供对该领域知识的共同理解，确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出这些词汇(术语)和词汇之间相互关系的明确定义。在特定的领域里，这样一个结构建立了很好的层次知识，对于一个网站，领域本体一般包括概念、概念之间的关系以及存在于网站表示的领域概念之间的关系。为了从网页或网站中抽取项目的语义信息，首要任务是建立一个特定领域本体。对于简单的网站，领域本体可能很容易手工建立或从网站内容半自动化获取.然而，对于大型网站，建造本体(特别是通用目的本体)是费时费力的过程，手工建立本体是一项艰巨的任务，希望能自动化地获取领域本体。文献[5,6]提出了 TextToOnto 系统和 OntoLearn 系统。综合起来，提出了一个本体学习的框架(见图 1)。

2)基于本体的抽取项目语义信息　建立了领域本体后，利用领域本体抽取项目的语义信息。在此方法中，使用特定领域包装器代理结合领域本体抽取项目的语义信息。特定领域包装器使用文本挖掘和启发式规则从基于领域本体的网站抽取项目的语义信息。目前，不使用本体表示语言，如 OWL，把本体中的类和类之间的关系看作一种关系数据库。特定领域包装器代理使用类的关系语法和基于文本线索的启发式规则抽取类的实例和属性，建立一个特定领域的语义分类树。以一个销售书籍网站为例，图 2 是从书籍网站抽取出

来的参考本体。从书籍网站抽取出来的实体都是这些类的实例, 这些实例包含了语义信息。

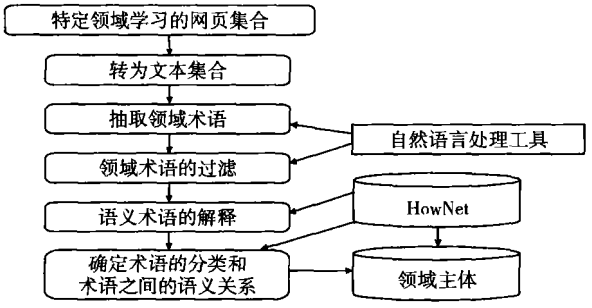

图1　本体学习框架

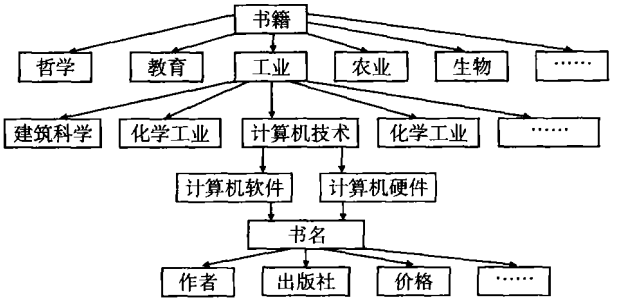

图2　书籍参考本体

3)语义相似性与协同过滤集成　为了方便计算项目语义相似性[7], 将抽取出来的实例属性转换为向量表示, 使用向量空间模型表示项目的属性[8], 项目可以表示为 $\boldsymbol{T}_i=\{(t_1, W_1), (t_2, W_2), \cdots, (t_i, W_i)\}$, 其中 t_i 是项目 $\boldsymbol{T}_i$ 的一个属性, W_i 表示属性 t_i 对应于项目赋予的权重, 描述属性在项目中的重要程度。项目之间的相似性使用向量之间的距离度量, 计算公式为

$$S(i, j)=\left(\sum_{k=1}^{M} W_{ik}\times W_{jk}\right)/\sqrt{\left(\sum_{k=1}^{M} W_{ik}^2\right)\left(\sum_{k=1}^{M} W_{jk}^2\right)} \tag{3}$$

采用 TF-IDF 公式计算属性的权重值[8]。其中 t_i 是项目 $\boldsymbol{T}_j$ 的一个属性。

$$W(t_i, \boldsymbol{T}_j)=tf(t_i, \boldsymbol{T}_j)\times \log(N/n_1+0.01)/\sqrt{\sum_{t_i\in T_j}[tf(t_i, \boldsymbol{T}_j)\times \log(N/n_1+0.01)]^2} \tag{4}$$

其中 $W(t_i, \boldsymbol{T}_j)$表示 t_i 在项目 $\boldsymbol{T}_i$ 的权重, 项目 $tf(t_i, \boldsymbol{T}_j)$是 t_i 在项目 $\boldsymbol{T}_j$ 中出现的频率, N 是项目总个数, n_1 为在项目集中出现 t_i 的数目。最终, 对于一对项目, 可以把语义相似性与协同过滤集成, 组合成一个线性方式来度量项目的相似性。

$$S_{\mathrm{Inte}}(i, j)=\alpha S_{\mathrm{Sem}}(\boldsymbol{T}_i, \boldsymbol{T}_j)+(1-\alpha)S_{\mathrm{Rat}}(i, j) \tag{5}$$

其中 $\alpha(0\leqslant\alpha\leqslant1)$是权重参数, 当 $\alpha=0$ 时, 组合的相似度 $S_{\mathrm{Inte}}(i, j)=S_{\mathrm{Rat}}(i, j)$; $\alpha=1$ 时, 组合的相似度 $S_{\mathrm{Inte}}(i, j)=S_{\mathrm{Sem}}(i, j)$。从公式(5)看出, 组合相似性算法具有 2 个优点: 组合相似性算法能进一步说明用户对特定的项目是否感兴趣; 在评分数据稀疏或没有评分的情况下, 依然可以使用语义相似性提供合理的推荐产品。利用该文组合相似性算法的思想, 可以得到相应的预测评分或推荐, $P_{a,i}$是用户对目标项目的预测评分值。$P_{a,i}$为

$$P_{a,i}=\sum_{j=1}^{k}(P_{a,j}\times S_{\mathrm{Inte}}(i, j))/\sum_{j=1}^{k} S_{\mathrm{Inte}}(i, j) \tag{6}$$

3　结果分析

1)数据集　采用一个销售书籍的网站数据来测试算法, 比较基于语义相似性和标准的项目协同过滤算法。该网站已有 1 600 个用户对 3 000 本书籍评分数据, 每个用户对每本书的评分范围为 1—5, 随机抽取 4 000条数据作为实验数据。为了实现基于语义相似性算法, 使用本体学习框架方法从网站书籍抽取出书籍本体, 利用包装器代理从基于书籍本体的网络书籍数据库中抽取书籍实例, 每个实例包含语义属性。

2)评价标准　评价推荐系统推荐质量的度量标准主要包括统计精度度量方法和决策支持精度度量方法 2 类。采用统计精度度量方法中的平均绝对偏差 MAE(mean absolute error)进行度量[3]。平均绝对偏差 MAE 通过计算预测的用户评分与实际的用户评分之间的偏差来度量预测的准确性, MEA 越小, 推荐质量越高。

假设预测的用户评分集合表示为$\{p_1, p_2, \cdots, p_n\}$, 对应的实际用户评分集$\{t_1, t_2, \cdots, t_n\}$, 则平均绝对偏差 MAE 定义为[3]

$$\mathrm{MAE}=\sum_{i=1}^{n}|t_i-p_i|/N \tag{7}$$

3)实验结果分析　考虑组合相似性算法中的 α 权重参数对 MAE 的影响, 在使用组合相似性算法预测

时,找出最优 α 值的范围。实验的结果见图 3,由此得出结论,α 的取值范围在 0.3—0.5 是最优的。为了说明组合相似性算法能解决标准项目协同算法的 2 个缺点:数据的稀疏性问题;新项目预测问题(冷开始问题)。做了 2 个实验比较这 2 种方法,2 个实验的 α 取值为 0.4。

图 4 显示了在预测项目都已评分的情况下 2 种算法的平均绝对偏差结果,从图 4 中可以看出,组合相似性算法使用公式(4)计算项目的相似性比标准项目协同算法使用公式(1)计算项目相似性更精确。

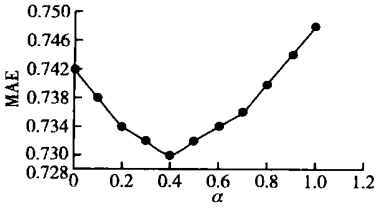

图3 α对MAE的影响

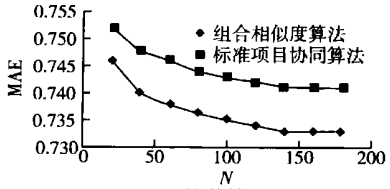

图4 推荐算法比较

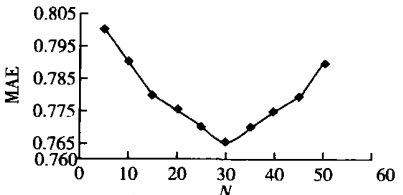

图5 基于组合相拟算法的新项目预测

图 5 解释了组合相似性算法可以解决新项目问题。在预测项目没有评分的情况下,基于项目协同过滤算法使用式(1)和式(2)无法预测,但组合相似性算法利用式(4)和式(5)依然可以进行预测。

从上述分析,通过语义相似性与标准项目协同算法集成,挖掘出项目之间的语义关系,抽取出项目的语义信息,不仅能很好的解决标准项目协同算法的项目评分的稀疏性问题、新项目预测问题以及提高推荐精度,还能进一步解释说明用户对特定的项目是否感兴趣。

4 结 语

通过集成项目的结构化语义信息计算项目的相似性,扩展了基于项目协同的过滤算法。通过本体学习建立了特定领域本体,并使用特定领域本体从网站中抽取项目的特性和聚集类的实例。相似性测量的方法是把基于领域的语义相似性与基于用户与项目映射的项目相似性组合。实验结果显示了组合相似性算法的 3 个优点:1)保持了基于项目协同算法的计算优势,组合相似性算法改善了预测精度。2)对于新项目或未评分的项目,能产生合理地精确的推荐,可以减轻标准项目协同过滤算法带来的新项目问题。3)在数据非常稀疏的情况下,提供了较高的预测质量。

下一步工作任务是:1)将使用领域特征和机器学习技术,实现自动地确定语义组合参数值。2)深入研究对语义相似性自动抽取以及度量。3)进一步考虑领域本体结构使用其他语义相似性组合方法产生推荐。

参考文献

[1] Breese J S, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering[A]. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence[C]. Madison: Morgan Kaufmann Publishers, 1998. 43-52.

[2] Resnick P, Iacovou N, Suchak M, et al. An Open Architecture for Collaborative Filtering of Netnews[A]. Proceedings of the ACM CSCW'94 Conference on Computer-supported Cooperative Work[C]. New York: ACM Press, 1994. 175-186.

[3] Sarwar B, Karypis G, Konstan J, et al. Item-based Collaborative Filtering Recommendation Algorithms[A]. Proceedings of the 10th Conference on World Wide Web[C]. Hong Kong: ACM Press, 2001. 285-295.

[4] Sarwar M, Karypis G, Konstan J, et al. Analysis of Recommender Algorithms for E-commerce[A]. Proceedings of the 2nd ACM E-commerce Conference (EC'00)[C]. Minneapolis: ACM Press, 2000. 158-167.

[5] Maedche A, Staab S. Learning Ontologies for the Semantic Web[A]. Semantic Web Worshop[C]. Hongkong: ACM Press, 2001. 72-79.

[6] Roberto Navigli, Paola Velardi. Ontology Learning and Its Application to Automated Terminology Translation[A]. [S. l.]: IEEE Intelligent Systems, 2003. 22-31.

[7] Vladimir Oleshchuk, Asle Pedersen. Ontology Based Semantic Similarity Comparison of Documents[A]. Proceedings of the 14th International Workshop on Database and Expert Systems Applications[C]. Prague: [s. n.], 2003. 735-738.

[8] Sun Yunheng, He Piliang, Chen Zhigang. An Improved Term Weighting Scheme for Vector Space Model[A]. Proceedings of 2004 International Conference on Machine Learning and Cybernetics[C]. [S. l.]: IEEE Systems, 2004. 1692-1695.