用户和项目联合度对二分网络个性化推荐的影响

程婷婷 王恒山 刘建国
(上海理工大学管理学院 上海 200093)

摘 要 首先采用物质流动算法进行二部图相似系数投影,然后利用随机游走模型得到协同过滤结果。在计算相似系数时,采用了考虑用户和项目联合度分布特征的改进算法。通过数据模拟可知,在最优情况下推荐项目准确率提高了18.19%,推荐项目多样性提高了21.90%。对用户和项目联合度的分布进行了统计分析,结果表明,在最优情况下,其符合指数为−2.33的指数分布。

关键词 个性化推荐,二分网络,协同过滤

Effort of User-Item Degree Correlations to Bipartite Network Personalized Recommendations

CHENG Ting-ting WANG Heng-shan LIU Jian-guo
(The University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract In this paper first bipartite graph was project based on mass diffusion, then random walk method was used to get collaborative filtering results. Degree correlation between users and objects was embedded into the similarity index to improve the algorithm. The numerical simulation shows that the algorithmic accuracy of the presented algorithm is improved by 18.19% in the optimal case and the diversity is improved by 21.90%. The statistical analysis on the product distribution of the user and object degrees indicates that, in the optimal case, the distribution obeys the power-law and the exponential is equal to −2.33.

Keywords Recommendation systems, Bipartite network, Collaborative filtering

1 引言

随着互联的扩展和 Web 2.0 技术的广泛应用,信息过载。怎样帮助用户高效获得其真正需要的信息,是今天面临的一个巨大挑战。个性化推荐通过历史数据预测用户的喜好和习惯,成为解决信息过载的一个有利工具,比如在 Amazon.com 推荐书和 CD、在 Netflix.com 推荐电影、在 Versifi Technologies(前身为 AdaptiveInfo.com)推荐新闻。由于个性推荐对商业和社会有着非常重要的现实意义,关于个性化系统的研究越来越多,在真实商业和数字图书馆互联网应用系统中推荐是关键组成部分。个性化推荐系统包括 3 部分:数据收集、模型分析和推荐方法。其中推荐方法是核心部分。个性化推荐方法包括 3 种:协同过滤推荐[1-4]、基于内容推荐[5,6]和混合推荐[9]。虽然个性化推荐被广泛研究,但一般只是关注数据集不变时推荐准确率的提高。真实的推荐系统由用户-项目二分网络组成,数据不停更新,所以算法的准确度并不能得到保证。为了得到一个更好的方法,本文从统计物理的角度研究系统数据的统计特性和算法结果的关系。

在过去的 10 年,协同过滤是其中较优且被广泛应用和研究的算法[1]。用协同过滤预测一个目标客户的潜在兴趣,首先根据历史记录计算目标用户与其邻居用户的相似系数,然后根据相似性加权采用邻居的历史选择项目。最近一些物理动力学过程,比如物质流动过程模型[10,11]和热传导过程[12],被创新地应用在协同过滤算法中。刘建国等[3]在计算用户之间的相似性时引入物质流动过程,通过计算标志算法的准确率的平均排名分数可知该改进算法较标准协同过滤算法准确率有大幅度提高;通过考虑高阶用户和项目关联矩阵,刘建国等[4]和周涛等[13]提出了一个准确率更高的算法,标志算法准确度的平均排名分数减小到约 0.08。以上两个算法都考虑消除由二阶用户相似关系矩阵或二阶项目相似关系矩阵带来的冗余。这两个算法并不能保证在其它训练集中的准确度,因为算法没有考虑训练集的统计特性。在以上两个算法中,具有不同度的用户和项目被同等对待,并没有考虑用户和项目的度。例如一个小度用户选择了一个小度项目,则代表这个用户明显具有非常独特的品味,该选择充分反映了用户的品味信息;反之,一个活跃用户选择一个流行项目,则从这个选择很难预测用户兴趣。因此,在计算相似系数时考虑二分网络中用户和项目度可以改进相似系数的准确性。数据模拟结果证明,降低主流偏好可以增加算法多样性和提高准确度。

2 构造协同过滤算法

在推荐系统中有两类结点,即用户和项目,用户选择项目

到稿日期:2010-06-12 返修日期:2010-09-20 本文受国家自然科学基金(71071098),上海市重点学科管理科学与工程(S30504)资助。
程婷婷(1981−),女,博士生,主要研究方向为系统工程、计算机应用技术;王恒山(1948−),男,教授,博士生导师,主要研究方向为数据挖掘、信息自动过滤算法、知识管理系统、决策支持系统;刘建国(1979−),男,硕士生导师,主要研究方向为信息过滤、复杂网络、数据挖掘、科学知识图谱分析。

的记录信息用代表用户的结点和代表项目的结点连线表示，这样就形成了用户-项目二部图。定义用户集合 $U=\{u_1,u_2,\cdots,u_N\}$，项目集合 $I=\{o_1,o_2,\cdots,o_M\}$，因此可以用 $N+M$ 个结点表示整个系统。这个二部图可以用一个连接矩阵来表示 $A=\{a_{i\alpha}\}\in R^{N,M}$。若用户 u_α 选择了项目 o_i，则 $a_{i\alpha}=1$，反之 $a_{i\alpha}=0$ 表示用户 u_α 没选择项目 o_i。在标准协同过滤算法中，第一步计算出用户之间或是项目之间的相似系数，第二步由各自的相似系数进行预测。如果用户 u_α 没选择过项目 o_i 即 $a_{i\alpha}=0$，则预测打分值为

$$score_{i\alpha}=\frac{\sum_{\beta=1}^{N}\mathrm{sim}_{\beta\alpha}*a_{i\beta}}{\sum_{\beta=1}^{N}\mathrm{sim}_{\beta\alpha}} \tag{1}$$

$\mathrm{sim}_{\mu\alpha}$ 表示用户 u_α 和用户 u_μ 之间的相关系数。其中两个计算相似度的经典公式为

$$\mathrm{sim}_{\beta\alpha}=\frac{2\sum_{h=1}^{M}a_{h\beta}*a_{h\alpha}}{d_{u_\alpha}+d_{u_\beta}} \tag{2}$$

$$\mathrm{sim}_{\beta\alpha}=\frac{\sum_{h=1}^{M}a_{h\beta}*a_{h\alpha}}{\sqrt{d_{u_\alpha}+d_{u_\beta}}} \tag{3}$$

式(2)称为 Sorensens 相似系数，由 Sorensens 在 1948 年提出。d_{u_α} 表示结点 u_α 的度。式(3)称为余弦相似系数，由 Salton 在 1983 年提出。对于一个目标用户 u_α，算法如下：

第一步　计算用户关联系数矩阵 $\mathrm{sim}_{\alpha\beta}\in R^{N,N}$；

第二步　对目标用户 u_α 根据式(1)预测没有选择过的项目的打分值；

第三步　把目标用户 u_α 没有选择过的项目的打分降序排列，推荐 top-θ(前 θ 名)。

3　用户和项目度之间的关系对协同过滤的影响

过去，相关系数只与用户的度、共同选择项目的数目有关，并没有考虑用户和项目度之间的关系。受物质流动过程启发，周涛等[11]和刘建国等[3]提出一种改进的协同过滤算法。此改进算法利用物质流动过程来计算用户之间的相似性，大大提高了算法准确率，同时证明多样性也提高了。新算法虽然改进了标准协同过滤算法，但并没有考虑用户度和项目度之间的关系，所以说所有边对于物质流动的贡献都是一样的。如果用户 u_α 和用户 u_β 同时选择一个项目 o_γ，那么他们可能有相同的兴趣和品味。假如 o_γ 的度非常大(即 o_γ 非常受欢迎)，这个喜好(喜欢 o_γ)非常普遍，并不能说明用户 u_α 和用户 u_β 的相似系数非常大。相反，假设一个小度用户 u_α 选择了一个小度(不流行的)项目 o_γ，这个兴趣或品味非常特殊，这个连接的贡献就应该扩大。如果一个大度用户选择一个大度项目，这个边对推荐没有多大意义。如果一个小度用户选择一个小度项目，这个边隐含了丰富的个性化喜好信息。所以用户 u_α 和项目 o_γ 之间的连线对个性化推荐的贡献反比于 $d_{u_\alpha}*d_{u_\beta}$。假设任意两用户之间存在某种数量的物质(也可以认为是推荐力)，权重 $\mathrm{sim}_{\alpha\beta}$ 表示用户 u_β 可能贡献给用户 u_α 的物质的比例。

因基于物质流动过程的网络结构推荐方法假定任一用户会把推荐能量分给其选择的项目，每个项目再把其收到的能量分给选择它的用户，所以考虑了用户-项目度关系的基于物质流动的用户相似系数计算公式为

$$\mathrm{sim}_{\alpha\beta}=\frac{1}{d_{u_\beta}}\sum_{h=1}^{M}\frac{a_{h\alpha}\cdot(d_{u_\beta}\cdot d_{o_h})^{\lambda}\cdot a_{h\beta}(d_{u_\alpha}\cdot d_{o_h})^{\lambda}}{d_{o_h}} \tag{4}$$

式中，λ 是控制度关系影响的调节系数。

4　基于二部图相似投影的随机游走推荐

由文献[2]物质流动模型可知，在达到一个平衡稳定状态后得到的物质(分数)才是最有效的。文献[2]中物质流动进行了两步，虽说大部分物质状态会收敛，但并不是太准确。本文根据文献[8]随机游走算法来解决这个问题，采用随机游走算法，当目标用户 u_α 对每一项目的打分向量 $score_{u_\alpha}=(score_{1\alpha},score_{2\alpha},\cdots,score_{M\alpha})'$ 收敛时，得到的结果较为准确。

算法步骤如下：

输入：表示用户与项目关系的二分网络 G，目标用户 u_α

输出：u_α 的推荐序列

1)根据改进相似系数计算式(6)得到 G 投影的一维用户关联矩阵(权重矩阵)$W=|\mathrm{sim}_{\alpha,\beta}|$。

2)把 W 对角线的元素设为 0，再以列为单位归一化，得到矩阵 W'。

3)取用户 u_α 对各项目打分的列向量 $score_{u_\alpha}=(score_{1\alpha},score_{2\alpha},\cdots,score_{M\alpha})'$，没有对项目打分的为 0，并将其归一化为向量 $score_{u_\alpha}^{stad}$。

4)$t=0$，初始化 $s_{u_\alpha}(t=0)=s_{u_\alpha}(0)=score_{u_\alpha}^{stad}$。

5)while(s_{u_α} 不收敛 ‖ $t<t_{\max}$)($t_{\max}$ 表示最大循环次数)

6){$s_{u_\alpha}(t+1)=W'*s_{u_\alpha}(t)$

$t++$ }

7)把 s_{u_α} 中的元素降序排列。

8)把前 top-θ 个项目推荐给用户。

5　算法优劣评价量

本文采用的训练集连线数占训练集和检测集总数的 80%。用平均排名分数测试算法的准确性，用 Hamming 距离测试算法推荐多样性，用平均度测试推荐是否流行。

5.1　平均排名分数

一个推荐算法会给每一用户提供一个降序排列的打分列表。对于一个随机用户 u_i，如果关系 u_i-o_j 在检验集里存在，计算 o_j 在降序列表中的位置。例如对于目标用户 u_i，列表长度为 $L_i=100$，S_j 在表中排第十位，称 S_j 的位置为 $10/L_i$，用 $r_{i,j}=0.1$ 表示。因为检验集中的连线是真实被用户选择的，一个好的算法希望在推荐列表中达到非常高的推荐度，所以要求 $r_{i,j}$ 较小。总的来说，所有检验集中所有连线的位置平均值，即平均排名分数 $\langle r\rangle$ 可以用来度量算法准确性：$\langle r\rangle$ 越小，准确度越高。

5.2　流行性和多样性

除了准确率，所有推荐项目的平均度 $\langle K\rangle$ 和平均 Hamming 距离 S 被引入，用来检验推荐的流行性和多样性。平均度越小，说明不是非常流行的项目也能被推荐，因为度小的项目少被人选，所以很少被用户在众多项目中发现。相异性可以由 Hamming 距离 S 来量化，$S=\langle H_{i,j}\rangle$，在这里 $H_{i,j}=1-Q_{i,j}/L$，因为推荐项目时把排名较大的项目推荐给用户是没有什么意义的，一般只推荐前 L 名项目，所以在计算 S 和 $\langle K\rangle$ 时只需计算前 L 支股票的平均值。$Q_{i,j}$ 代表用户 S_i 和用户 S_j 长度为 L 的推荐列表相同推荐的数目。S 最大值为 1，代表所有用户项目推荐列表全不相同。当 $S=0$ 时，代表所有用户的推荐列表完全相同。

6 模拟数据结果

本文把上面的算法应用于 MovieLens 数据，该数据包括 943 个用户和 1682 部电影。数据有 3 项：用户 u、电影 o、打分 $score$($score\in\{1,2,3,4,5\}$)。本文对数据粗粒化处理，即只保留打分大于 2 的数据，并且把打分粗粒化为 1。原始数据包括 10^5 个打分，粗粒化后数据有 85520 个打分。

表 1 各种算法标志量对比

Algorithms	⟨r⟩	S	⟨k⟩
GRM	0.1390	0.398	259
CF	0.1168	0.549	246
ICF	0.1156	0.630	229
ZhouM	0.0820	0.793	175
MCF	0.0998	0.692	218

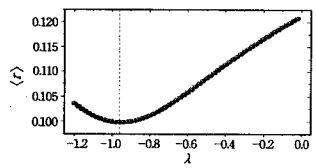

图 1 ⟨r⟩关于参数 λ 的函数曲线

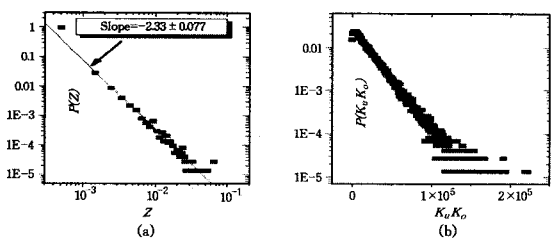

图 2 联合度分布，图(a)$\lambda^{opt}=-0.96$，$z=(d_u*d_o)^{\lambda^{opt}}$；(b)$\lambda=1$

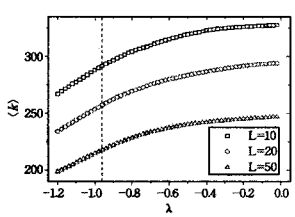

图 3 平均度关于 λ 的函数曲线

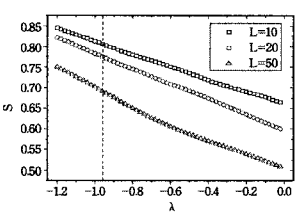

图 4 S 关与 λ 的函数曲线

图 1 是⟨r⟩关于参数 λ 的函数曲线，很明显曲线在 $\lambda^{opt}\approx-0.96$ 时⟨r⟩有最小值，这充分验证了降低度大的用户和项目对相似系数的影响能提高算法的准确度。相对于未改进时算法(即 $\lambda=0$ 时)，⟨r⟩在最优值减少了 18.2%，明显提高了算法的准确度。表 1 是几种算法结果标志量列表，本文算法 MCF(Modified collaborative filtering)远优于 GRM(总体排名推荐)和 CF(标准协同推荐)，却不如 ZhouM。表中 ICF[3] 和 ZhouM[13] 算法的数据分别对应用于参数 $\lambda_{opt}=1.9$ 和 $\beta_{opt}=-0.8$ 的情况。ZhouM 方法考虑用二阶关系矩阵消除关系冗余，大大提高了算法准确度，但也因此增加了算法时间复杂度，使算法较难应用推广。图 2 是关于用户和电影联合度的分布图，图 2(a)($\lambda^{opt}\approx-0.96$)显示不同于图 2(b)($\lambda=-1$)，联合度分布符合指数为 -2.33 ± 0.077 的指数分布。图 3 是平均度关于 λ 的函数曲线，$\lambda<0$ 时⟨K⟩随 λ 的增大单调增加，因此证明消弱大度用户与大度电影连接边对关联系数的影响，大大增加了非流行电影得到推荐的机会，这正是算法希望达到的目的之一。图 4 显示 S 与 λ 成反比关系，这证明消弱大度用户与大度电影连接边对关联系数的影响，算法得到的推荐电影更加个性化。对于 $L=10$，当 $\lambda=0$ 时 $S=0.661$，当 $\lambda^{opt}=-0.96$ 时 $S=0.806$，易知个性化程度提高了 21.90%。

结束语 本文提出了考虑用户和项目联合度的基于物质分配网络结构协同推荐的改进算法。数据模拟结果验证降低活跃用户和流行电影连接在推荐中的权重，通过计算平均排名分数可知算法准确度提高了 18.2%。$\lambda^{opt}=-0.96$ 时，度小的用户和度小的电影之间的连接的影响被加强了，并且联合度分布符合指数形式的分布。虽然不能清晰地解释指数分布和最优准确值之间的关系，但事实证明这两者存在一定的数学关系。除准确率外，另外两个检测算法优劣的重要准则是多样性和流行性。一个好的算法，不但应该具有较高的准确性，而且能帮助用户发掘隐性信息(对应于度小的项目信息)。所以推荐项目的平均度是评定个性化推荐算法的一个有重要意义的指标。另外，个性化推荐应该根据用户习惯和兴趣提供个性推荐，所以推荐的相异性是检测个性化程度的重要量。数据模拟结果证明本文提出的算法在准确率、流行性和多样性 3 个方面都优于标准协同过滤算法(CF)。

从数据模拟结果易知，λ^{opt} 近似 -1。当 $\lambda=-1$ 时，⟨r⟩$=0.0995$，准确率提高了 18.2%，平均度和多样性也得到了明显改善，比如多样性 S 改进了 23.4%。所以在真实应用中，把 λ 值设为 -1，这样增加参数 λ，时间复杂度并没有增加。考虑用户和项目度关系的改进算法可以提高算法的准确率，时间复杂度也比标准 CF 更小。但怎样在不同数据集中找到 λ^{opt} 制约了算法的应用。对于实验数据至少有两种方法去寻找 λ^{opt}：第一种方法首先在一个小数据集上试验，寻找系数 λ 与算法结果优劣的关系；第二种分析用户和项目联合度的分布。本文发现最优值联合度分布满足指数分布，虽然现在还不能完全解释其中的联系，但指数和最优值之间存在一定的内在关系，这将是改进算法的一个新角度。

将来的工作将更加关注结构属性(比如聚类系数 C_3 和 C_4)与算法表现的关系。怎样自动找出不同用户的相关信息，是现代信息科学的一个长期挑战。这个方法也可以应用于查找相关的科技论文和专利网中，预测社会和生物网的链接。相信本文会给读者提供一个新的研究角度。

参 考 文 献

[1] Herlocker J L, et al. Evaluating Collaborative Filtering Recommender Systems[J]. ACM Trans, Inform, Syst, 2004, 22(5)

[2] Konstan J A, et al. Applying collaborative filtering to usenet news [J]. Commun. ACM, 1997, 40: 77

[3] Liu J-G, Wang B-H, Guo Q. Imroved collaboration filtering algorithm via Information transformation [J]. Int. J. Mod. Phys, 2009, C 20: 285

[4] Liu J-G, et al. Effects of high-order correlations on personalized recommendations for bipartite networks[J]. Physica A, 2010, 389: 881

[5] Balabanovś M, Shoham Y. Content-based, Collaborative Recommendation [J]. Commun, ACM, 1997, 40: 66

[6] Pazzani M J. A Framework for Collaborative, Content-based and Demographic Filtering [J]. Artif. Intell. Rev., 1999, 13: 393

[7] Ou Q, Jin Y-D, Zhou T, et al. Power-law Strength-Degree Correlation From a Resource-Allocation Dynamics on Weighted Networks [J]. Phys. Rev. E, 2007, 75: 021102

[8] Gori M, Pucol A. A random-walk based scoring algorithm with application to recommender systems for large-scale ecommerce [C]// Proceedings of WEBKDD'06. Philadelphia, USA, 2006: 127-146

[9] Pazzani M, Billsus D. Learning and revising user profiles: The identification of interesting web sites[J]. Machine Learning, 1997, 27: 313

[10] Zhang Y-C, et al. Recommendation model based on opinion diffusion [J]. Europhys. Lett., 2008, 80: 68003

[11] Zhou T, et al. Bipartite network projection and personal recommendation [J]. Phys. Rev. E, 2007, 76: 046115

[12] Zhang Y-C, Blattner M, Yu Y-K. Heat Conduction Process on Community Networks as a Recommendation Model [J]. Phys. Rev. Lett., 2007, 99: 154301

[13] Zhou T, et al. Accurate and diverse recommendations via eliminating redundant correlations[J]. New J. Phys, 2009, 11: 123008

(上接第 177 页)

用户访问序列经过 HMM 处理后的用户兴趣类别序列作为序列模式挖掘的数据源，存在数据库中，形成序列数据库，共包含 30 个序列。GSP 算法运行时的最小支持度可根据实际情况设置，支持度的大小会直接影响到算法执行的时间和挖掘出的模式。例如，如果模式级别定义得比较高，那么在预处理阶段就有可能清洗掉很多低级类别的用户兴趣浏览记录。此时如果将最小支持度定义得比较大，就很难发现长序列模式。算法运行时所需数据直接从序列数据库中获取。将最小支持度分别设置为 3，4，5 时，算法执行结果如表 2 所列。

表 2 在不同的支持度阈值下 GSP 算法执行的结果

最小支持度=3		最小支持度=4		最小支持度=5	
用户兴趣迁移模式	模式支持度	用户兴趣迁移模式	模式支持度	用户兴趣迁移模式	模式支持度
军事→经济→环境→计算机	4	军事→经济→环境	7	军事→经济→环境	7
		经济→军事→计算机	4		
		经济→环境→计算机	4		

本文在测试时把最小支持度设置为 3，程序运行时间为 62ms，发现的序列模式很多，最长的序列模式长度为 4。下面以长度为 4 的序列模式〈军事、经济、环境、计算机〉为例解释该模式的意义：该模式表示在所有用户访问兴趣序列中，至少有 4 天的时间用户在 Web 上都有这样一个访问路线：军事→经济→环境→计算机。可以看出用户对这 4 个类比较感兴趣。

序列模式挖掘出来的结果是从全局的角度挖掘出的用户兴趣类别，反映了一个用户具有持久的兴趣倾向和用户访问 Web 的偏好。通过分析可知，最长序列模式中包含的兴趣类别最能反映用户访问兴趣的稳定性。本系统中将挖掘出的最长序列模式中的兴趣类别作为用户兴趣模型中的稳定兴趣。在不同支持度阈值的设定下提取的用户稳定兴趣结果如表 3 所列。

表 3 用户稳定兴趣提取结果

最小支持度=3		最小支持度=4		最小支持度=5	
用户兴趣迁移模式	用户稳定兴趣	用户兴趣迁移模式	用户稳定兴趣	用户兴趣迁移模式	用户稳定兴趣
军事→经济→环境→计算机	军事、经济、环境、计算机	军事→经济→环境 经济→军事→计算机 经济→环境→计算机	军事、经济、环境、计算机	军事→经济→环境	军事、经济、环境

从表中可以看出，当算法中最小支持度为 3 时用户稳定兴趣为军事、经济、环境、计算机；算法中最小支持度为 4 时用户稳定兴趣为军事、经济、环境、计算机；算法中最小支持度为 5 时用户稳定兴趣为军事、经济、环境。

为了评估本文中提出的兴趣迁移方法的性能，定义算法的查全率和精确度：

查全率＝正确识别出用户的兴趣数/专家推荐的用户兴趣总数；

精确度＝系统正确标记的用户的兴趣类别/系统标记的用户兴趣类别总数。

考虑到专家的知识是主观的，不能提供精确的定量分析，我们邀请了 5 组专家来做同一个试验，以期得到一个比较客观的结果。经过分析，在本实验中，专家推荐的用户兴趣总数最终定为 5，即计算机类、体育类、环境类、经济类、军事类。

实验中，当把支持度阈值设为 3 时，提取出了 4 个稳定兴趣，算法的查全率为 0.8；支持度阈值设为 4 时，提取出了 4 个稳定兴趣，算法的查全率为 0.8；支持度阈值设为 5 时，提取出了 3 个稳定兴趣，算法的查全率为 0.6。

结束语 个性化 Web 服务通过收集和分析用户信息来学习用户的兴趣和行为，从而达到主动推荐的目的。个性化服务技术能充分提高站点的服务质量和访问效率，从而吸引更多的访问者。实现个性化服务的关键就是正确分析 Web 用户浏览信息，准确地描述用户的兴趣。只有准确地把握用户的浏览兴趣以及兴趣的变化情况，才能将用户感兴趣的资源推荐给用户，更好地描述用户兴趣。

参 考 文 献

[1] Land C, Soltysiak S. Identifying and Tracking Changing Interests[J]. International Journal of Digital Libraries, 1998, 2: 38-53

[2] Cxand W, Kubat M. Learning in the Presence of Concept Drift and Hidden Contexts[J]. Machine Learning, 1996, 23: 69-101

[3] Maloof M, Michalski S. Selecting Examples for Partial Memory Learning[J]. Machine Learning, 2000, 41: 27-52

[4] Koychev I, Schwab I. Adaptation to Drifting User's Interests [C]// Proceedings of ECML2000 Workshop: Machine Learning in New Information Age. Barcelona, Spain, 2000: 36-45

[5] Ding Y, Li X. Time weight collaborative filtering [C]// Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM). New York, NY, USA: ACM Press, 2005: 485-492

[6] Rabiner L R. A Tutorial on Hidden Markov Models and selected Applications in Speech Recognition [J]. Proceeding of the IEEE, 1989, 77: 257-286

[7] 刘立军，崔杰，梅红岩. GSP 与 PrefixSpan 算法的比较与分析[J]. 辽宁工学院学报，2006(05)

[8] 谭薇. 基于 Web 访问信息的用户兴趣迁移模式的研究[D]. 西安：西安邮电学院，2010

[9] 宗成庆. 统计自然语言处理[M]. 北京：清华大学出版社，2008

[10] 刘河生，高小榕，杨福生. 隐马尔可夫模型的原理与实现[J]. 国外医学：生物医学工程分册，2002, 25(6): 253-259