信息协同过滤*

蔡登，卢增祥，李衍达
清华大学 自动化系 北京 100084
caideng00@mails.tsinghua.edu.cn
luzx@tsinghua.edu.cn

摘要：

Internet 的迅速发展使得信息过载日益严重，如何帮助用户有效的发掘、过滤、利用信息这一问题亟待解决。在各种信息过滤算法中，协同过滤以其独有的特点得以迅速发展。本文试图描述协同过滤的本质，分析阐述实际中和实验中的一些算法，探讨协同过滤的发展方向。

关键词：信息过滤 协同过滤 社会过滤

1. 简介

网络的迅速发展、信息的日益丰富使得信息过滤越来越重要。在网络发展初期发挥了重大作用的搜索引擎正面临着困境：网络资源的众多和低组织性使得搜索引擎无法准确的根据用户提交的查询返回用户需要的内容。单一的关键词提供的信息量太少，难以据此准确判断用户的需求。而传统的基于内容（关键词匹配）的过滤技术在人工智能自然语言理解没有重大突破之前已很难再有较大的发展，难以满足用户的进一步需求。

在这种背景下，协同过滤（Collaborative Filtering）的出现为进一步提高信息服务质量提供了一个新的思路。协同过滤，又称社会过滤（Social Filtering），其基本思想十分直观：在日常生活中，人们往往会根据亲朋好友的推荐来做出一些选择（购物、阅读、音乐……）。协同过滤系统就是将这一思想运用到网络信息服务（信息推荐）中，基于其他用户对某一信息的评价来向某一用户进行推荐。通常，系统选取与指定用户有相似兴趣的用户作为参考对象。而如何定义用户相似性以及如何选取参考用户群正是协同过滤算法研究的重点。

与传统文本过滤相比，协同过滤有下列优点：

1. 能够过滤难以进行机器自动内容分析（Content_based）的信息。像艺术品、音乐。
2. 共享其他人的经验，避免了内容分析的不完全和不精确，并且能够基于一些复杂的，难以表述的概念（如信息质量、品味）进行过滤。
3. 可以有效的使用其他相似用户的反馈信息，减少用户的反馈量，加快个性化学习的速度。
4. 具有推荐新信息的能力 （serendipitous recommendations）。

正因为此，在 Goldberg 等人[1]在其设计的邮件过滤系统中初步应用了协同过滤的思想

*基金项目：国家自然科学基金（60003004）。**蔡登**，硕士生，研究方向为网络信息服务，人工智能，MAS。**卢增祥**，博士，讲师，研究方向为网络信息服务，人工智能，MAS。**李衍达**，教授，博士生导师，中国科学院院士，研究方向为网络信息服务，生物信息学，智能信号处理等。

（这可以说是最早的协同过滤系统）之后，各种研究协同过滤的实验系统纷纷出现。像 GroupLens[2-4]：过滤网上新闻的系统；Ringo[5]：推荐音乐的系统；Video Recommender[6] 和 MovieLens[7]：推荐电影的系统；Jester[8]：推荐笑话的系统。

越来越多的在线商家，包括 Amazon.com、CDNow.com 和 Levis.com，都使用了协同过滤技术向顾客推荐产品[9]。由微软研究院开发的协同过滤工具已被集成在微软的 Commerce Server[1]产品中，并被许多站点使用了。

当然，协同过滤还有很大的不足，这是由协同过滤的本质决定：要获得满意的效果，必须得到准确的用户信息，一般而言，这需要建立在拥有大量的用户信息数据的基础上，所以很难做到。这使得协同过滤技术应用领域较为狭窄（几乎都集中在娱乐方面：音乐、电影……），在更广的领域（如在基于内容过滤相当成功的文本相关性领域）的应用还很不够。解决的途径不外乎两条：

1．发展信息获取技术：利用各种途径，合理有效的获得、使用更多的用户信息。

2．发展信息挖掘技术：在有限的原始数据中最大程度的找出有用的信息。

本文以这两条思路为线索，试图通过回顾协同过滤发展的历程，分析阐述实际中和实验中的一些算法，来揭示协同过滤的本质，并探讨协同过滤的发展方向。

2. 协同过滤的实现

协同过滤的基本出发点是：

1. 用户是可以按兴趣分类的；
2. 用户对不同的信息评价包含了用户的兴趣信息；
3. 用户对一未知信息的评价将和其相似（兴趣）用户的评价相似。

这三条构成了协同过滤系统的基础。

协同过滤的实现一般分为两步：首先，获得用户信息（用户对某些信息条目的评价……）；其次，分析用户之间相似度，预测特定用户对某一信息的喜好。上文提到的解决协同过滤不足的两条途径正是对应着这两步。

2.1 用户信息的获取

用户信息的获取主要通过用户对给定信息的评价。

评价分为显式评价（explicit rating）和隐式评价（implicit rating）两种。

显式评价需要用户有意识的表达自己对某一信息的认同程度，一般用整数值来表示喜欢的不同程度，如 GroupLens[2, 3]、Ringo[5]……，协同过滤系统向新用户提供一个信息列表，要求用户对其中全部或部分信息进行评价，系统获得用户的这些初始信息后，就能将用户加入到用户库中，随着用户不断使用协同过滤系统，用户的信息不断积累。

隐式评价希望从用户的行为中获得用户信息，目前已作的研究有通过分析用户网上购物记录、阅读文章的时间[10]、URL 的连接次数[11]等数据记录获取。

就目前而言，由于隐式评价对数据的分析难度较大，准确性、有效性有待于进一步提高。所以协同过滤领域的研究主要以显式评价为主。下文对于协同过滤算法的分析主要基于显式评价。然而从发展潜力看，隐式评价由于其数据量的几乎无限性，随着数据分析准确性的提高，将会占越来越大的比重。

[1] http://www.microsoft.com/directaccess/products/net/commerce/

2.2 协同过滤算法

对于协同过滤一个直观的描述是：将用户和信息条目构成一个矩阵：用户—信息条目的兴趣矩阵。（图 1）

	泰坦尼克	星球大战	007	谍中谍
张三	?	2	4	4
李四	4	?	2	2
王五	4	3	?	?
赵六	4	3	2	?

图 1

矩阵中已有的值是用户对相应信息条目的评价，未知值正是需要系统给出的预测。协同过滤的过程就是根据已知值来预测未知值（一个填空过程）。协同过滤系统所应用的算法就是这一填空过程所遵循的规则，规则与实际规律越符合，预测的未知值就越准，信息过滤的效果就会越好。需要注意的是，实际中，这样的矩阵是一个极为稀疏的矩阵。因为每一个用户一般只会对所有信息条目中很少一部分有评价。这一点在设计算法，分析算法优劣上都很重要。

已有的协同过滤算法可以分为两个方向[12]：

1. 全局数值算法（Memory-based algorithms）在对某个特定用户作预测时，需要对整个用户数据库进行比较、计算。如 GroupLens[2, 3]、Ringo[5]采用的算法。
2. 基于模型算法（Model-based algorithms）利用用户数据估计、学习一个用户模型，根据模型作出预测。

前者通过选取一个相似用户集，综合这个相似用户集中用户对一给定信息条目的评价，形成对给定用户的预测；后者一般采用概率模型，在已有的信息基础上，估计给定用户对信息条目的评价分布，进而得到期望，作为算法的预测值。

不管是哪一种方法，其最终目标就是上文说的途径二，即在有限的原始数据（已知的用户评价）中最大程度的找出有用的信息（用户之间的相关性）。

2.2.1 全局数值算法（Memory-based algorithms）

全局数值算法通过选取一个相似用户集，综合这个相似用户集中用户对一给定信息条目的评价，形成对给定用户的预测。直观的，这一过程可以分为三步：首先，定义一个距离准则，用于衡量用户之间的相似度；其次，根据用户之间的相似性选取相似用户集；最后，按特定的方法，综合相似用户集中用户对给定信息条目的评价，形成系统对选取用户的预测。

已有算法的区别主要在第一步，采用不同的相似性度量准则，后两步上的努力则是对基本算法的改进，以达到更好的效果。

目前相似性度量准则主要有：皮尔森相关系数（Pearson correlation coefficient）[2]、有约束的皮尔森相关系数（constrained Pearson correlation coefficient）[5]、向量相似性[12]，以及在这些准则基础上的一些变体。其中，皮尔森相关系数在协同过滤系统中使用得最为广泛，实验[4, 12]也证明了它得有效性。皮尔森相关系数的基本公式如下：

$$k_{xy}=\frac{\sum_{j\in I_{xy}}(r_{xj}-\bar{r}_x)(r_{yj}-\bar{r}_y)}{\sqrt{\sum_{j\in I_{xy}}(r_{xj}-\bar{r}_x)^2}\sqrt{\sum_{j\in I_{xy}}(r_{yj}-\bar{r}_y)^2}} \tag{1}$$

其中k_{xy}即为用户 x 与 y 之间的皮尔森相关系数，用于表征两用户之间的相似程度；r_{xj}是用户 x 对 j 信息的评价值；$\bar{r}_x$是用户 x 对所有信息的评价均值；I_{xy}代表 x 和 y 都评价了的信息集。

在第二步中，可以直观猜想，选取一个用户子集而不是所有用户来计算预测值会更准确。因为我们排除了与指定用户不相似的用户，相当于除去了噪声。实验[5]也证实了这一点。在如何选择用户子集上，通常的做法是定一个相似度的阈值，选择相似度大于阈值的用户；或者按相似度大小选择确定数量的用户；还有将这两种方法结合在一起。

综合用户评价通常采用如下公式：

$$p_{xj}=\bar{r}_x+\Lambda\sum_{y=1}^{n}k_{xy}(r_{yj}-\bar{r}_y) \tag{2}$$

其中，p_{xj}为系统预测 x 用户对信息 j 的评价；Λ是规范因子，可以是所有k_{xy}之和的倒数。求和号对选择的用户子集作用。其余符号含义同（1）处。

全局数值算法，主要是基于相关性的方法（correlation-based）在信息充分（用户评价矩阵只有少量空缺值）的情况下，有相当不错的效果。然而，正如前面提到的，实际问题中用户信息往往不足，需要设计能利用更多信息，能更有效利用信息的算法。这在下面详细阐述。

2.2.2 基于模型算法（Model-based collaborative filtering）

基于模型算法，是从概率的角度看协同过滤：在给定所有信息条件下，计算（估计）特定用户某信息条目评价的期望。

假设用户评价值是 0 到 m 的整数，对于给定用户 a，取 a 对信息 j 的评价

$$E(r_{a,j})=\sum_{i=0}^{m}P(r_{a,j}=i\mid A)i \tag{3}$$

表达式中的条件概率指在已知评价矩阵 A 的条件下，用户 a 对信息 j 评价值为 i 的概率。如何计算这一条件概率即是基于模型协同过滤算法的重点。

通常的做法是将用户分成 n 类 C_i（i =1~n），由已知用户数据估计出每类对信息条目 j 的评价分布$P(r_{c,j}=i\mid c)$，再由选定用户的已知信息估计出该用户属于某类的概率$P(a\in c\mid r_{a,k},k\in I_a)$，结合这两者最终得到选定用户对信息条目 j 的评价。这里 C 类即是用户的 C 个模型。

在基于模型的各种方法中，Pennock 等[13]的“个性诊断”算法较为突出：

Pennock 将每个用户模型描述成该用户对所有信息条目真实喜好值构成的向量，而用户

给出的评价值是在其真实喜好值上加一高斯噪声得到。将每一不同的用户模型作为一类，应用上面描述的算法即可得到结果。

可以看到，这里所说的模型只是概率模型，并不涉及具体的用户建模，使用的数据也仅限于用户的评价数据。能否进一步扩展用户模型？将在下面详细阐述。

3．协同过滤的发展方向

分析了协同过滤结构及其基本算法之后，我们可以看到用户信息在其中的决定性作用，只有在足够的用户信息条件下，全局数值算法才能准确的找到相似用户群，基于模型算法才能准确的估计模型参数。

要解决用户信息不足这一问题，已有的工作是从两个方向努力的：

1． 在有限的原始数据中最大程度的找出有用的信息。

2． 利用各种途径，合理有效的获得、使用更多的用户信息。

建立有效的用户模型也许是另一条出路。当然，要真正提高信息服务质量，协同过滤必需和其它信息技术相结合。

3.1 信息挖掘方向的努力

现有的协同过滤系统一般都采用用户对信息条目的显式评价作为初始信息。这一初始信息通常表示为用户—信息条目评价矩阵。如何从这一评价矩阵中获得更深层次的用户相关信息便是在这一方向上努力的目标。

在上文描述的全局数值算法中，用户之间相似性的计算必须基于用户对相同条目做出了评价，也就是说不同用户评价了的条目之间必须有重叠部分；而且要得到较为准确的预测，重叠部分条目数目不能太少，否则一两项内容的评价值将会决定用户之间的相似性。然而，用户—信息条目的兴趣矩阵是一个稀疏矩阵，这样，基于相关系数的算法使用范围受到很大限制，很多情况下不能很好的表达出用户的相似性。更主要的是采用该算法的系统对于一个与其他用户没有共同条目的新用户将无法作出预测，这在协同过滤系统的初始阶段（用户量较少时）尤为明显。

而直观认识告诉我们，当两个用户之间没有公共评价信息条目，但两者都与第三者有公共条目时，还是可以得到一些这两个用户的相似信息的。譬如用户 A 和用户 B 有很高的相似性，B 和 C 也有很高的相似性，这表明 A 和 C 之间也会有一定相似性，但如果在用户—信息条目兴趣矩阵中 A 和 C 没有公共评价条目，采用上文所述的算法就无法得到这一结果。所以，采用直接计算两用户相关系数的邻居选择方法丢失了一些原始信息。如果有效的利用这些信息，就可以扩展协同过滤应用面，更进一步也许可以提高预测准确度。

从信息条目的角度看，信息的类似反映在评价矩阵中即是信息对应的行向量相关（信息条目的相似导致用户对其评价值一致）。这样就可以对评价矩阵压缩降维，发掘其中信息的内在联系

这一思想可以与传统的 IR（Information Retrieval）中的隐含语义索引（LSI）[14-16]类比。这样，奇异值分解（SVD）这一能够降低评价矩阵维数并且发掘数据潜在联系的有效工具就可以被应用在协同过滤中。已有的工作已证明了其有效性。[17-19]

3.2 信息获取方向的努力

在挖掘已有数据内部规律的同时，有效合理的获取更多用户信息也许是一条更有效的途径。

协同过滤的本质是根据用户对信息的评价得到用户的兴趣，判断相似用户，从而向用户推荐新信息。在这一过程中，系统作了两次推断：从用户对一些信息的评价推断出用户的兴趣；从用户的兴趣推断出相似用户群。已有的协同过滤系统在这两次推断中所用的信息通常为用户对一些信息条目的显式评价。事实上，在这两步推断中，都还有很多其他信息可以帮助完成目标。

在第一步推断中，如前所说，除了显式评价，用户的兴趣还能从更丰富的隐式评价中获得。在利用隐式评价过程中，需要解决的问题有：有哪些隐式评价？可以分成几类？隐式评价的有效性如何？如何将隐式评价信息和显式评价信息有效结合？现有的工作[20, 21]按用户的行为对隐式评价进行分类，譬如“检查”代表用户阅读了整个信息；“保存”表示用户保存了信息，或将该信息加入了书签，或打印了信息……。另一些工作[3, 10, 22]对隐式评价的有效性进行了研究，实验证明了一些隐式评价（如阅读时间）对反映用户兴趣的有效性。对隐式评价的工作还处于开始阶段，上面提到的问题还没有很好的解答。

在第二步推断中，判断两用户的相似，也有一些其它信息可以利用：譬如两用户的职业相同，他们通常都会对某些信息感兴趣。利用用户的这些非评价信息将有可能更好的确定相似用户。Shapira 和 Shoval [23]在其设计的信息过滤系统中初步利用了这些信息，更深入的研究还有待进行。

3.3 建立有效的用户模型

上面已经指出：现有的基于模型算法并不涉及对用户具体描述建模。但如果我们在已有的信息基础上建立对用户的描述（profile），以此建立用户模型，然后根据模型找相似用户，进行协同过滤。这种方法可能的优点：

1. 协同过滤重要的是找到用户之间的一致性信息，但是由于用户评价资源不能很多，造成信息稀疏，使协同过滤无法进行，利用用户模型的方法能够压缩信息空间，这样，在用户模型之间寻找一致性信息将比较容易。
2. 已有的协同过滤算法（全局数值算法），每次预测都要计算所有用户之间相似度，而随着用户数据库的增大，信息条目的增多，计算所有用户之间相似度资源消耗太大。利用用户模型，就有可能根据用户反馈信息逐步调整模型，避免每次推荐重新计算。

用户模型一般分为通用模型和个人模型。前者又称为固定规则（Stereotype Rules）[24]，是对一组用户有效的知识。后者又称为用户描述（User Profile），是该用户特有的信息。当然，二者并非截然不同，协同过滤正是要根据用户模型在不同用户间确定相似度。

用户模型的描述一般有两种方式：

1. 知识库的表示方法。记录用户的职业、爱好等知识。
2. 间接的表达方式，例如由一组关键词（带权重）组成。或者使用用户评价后的文章来描述用户的兴趣。

采用第一种方式，可以实现一定程度的协同过滤。文献[23]根据用户的社会属性（职业、学历……）确定相似度，实现了协同过滤。但是协同过滤的优势领域（基于一些复杂的，难以表述的概念进行过滤）难以用此方法表示。第二种方式似乎是更好的解决方法，但是如果只停留在这个基础上，事实上又回到了协同过滤全局数值算法上。

用户建模是人工智能领域另一重要课题，目前也处于研究阶段。借鉴这一领域的技术，协同过滤系统的设计也许可以开辟出一条新路。

3.4 与其它信息技术的结合

作为信息过滤的一种方法，以向用户提供高质量的信息服务为目标，协同过滤技术只有和其它的信息过滤、信息推荐方法结合起来，才会发挥更大的作用。

传统的基于内容（content-based）的信息过滤技术发掘了信息内容这一方面信息；协同过滤则发掘了另一方面信息（用户之间的联系）。直观推想，将两者结合起来能取得更好的效果。但如何做呢？基于内容的算法没有提供一条明显的利用其他用户信息的途径；协同过滤算法也没有提供一条明显的利用信息内容的途径。已有的研究在这一方向做了一些尝试，如：Basu，Hirsh[25]将 Ripper[26, 27] ，一个诱导规则学习系统（inductive learning system）使用在协同过滤中，同时结合了文本过滤。它也是将问题看作一个分类过程，以集合属性（包括用户相似属性和信息文本属性）描述样本，进行分类。还有 WebWatcher，一个结合内容分析和协同过滤技术学习用户网页浏览习惯的个人助理[28]。

多代理系统（MAS）是目前信息过滤领域另一个热门话题。理想中的 MAS，用户有代表个人的助理（Personal Agent），能够主动与其它 Agent 协作，完成任务。协同过滤完全是可以与之结合的，个人助理（Personal Agent）正是一个有效的用户模型。当然，协同过滤在 MAS 中的应用有其新的特点：目前的协同过滤系统是集中式的，用户之间的相似度可以统一计算，而 MAS 是分布式的，如何形成相似用户群呢？已有的一些想法有：利用信任度，个人代理之间通过交互，不断积累对方的信息，形成对对方的信任度（相当于相似度），从而形成相似用户群[29]。

在自动协同过滤系统 ACF(Automated collaborative filtering)的实践中，另一些研究[30]是借鉴专家系统中人机交互的成功实例，认为在 ACF 系统中也应有人机交互（can be questioned）：向用户提供一个推荐过程直观的解释，使用户参与推荐决策，获得更好的推荐效果。

4．总结

协同过滤（Collaborative Filtering）这一概念的正式提出[1]到现在大约才十年，较之传统的文本分析方法突出的优点使其迅速发展。就协同过滤本身的发展而言，我们可以从信息获取和信息挖掘这两条路入手设计更完善的算法；从信息过滤整个问题角度看，我们应该将协同过滤与其它过滤技术（如文本过滤），与其它好的思想（如分布式代理）结合起来，这样才能更充分的发挥其作用。

参考文献：

[1] Goldberg, D., Nichols, D., Oki, B. M., and Terry, D., "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, Vol. 35, No. 12, 1992, pp. 61-70.

[2] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J., "GroupLens: An open architecture for collaborative filtering of netnews," In *Proceedings of 1994 Conference on Computer Supported Collaborative Work*, 1994, pp. 175-186.

[3] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., and Riedl, J., "GroupLens: Applying collaborative filtering to Usenet news," *Communications of the ACM*, Vol. 40, No. 3, 1997, pp. 77-87.

[4] Herlocker, J. L., Konstan, J. A., Borchers, A., and Riedl, J., "An algorithmic framework for performing collaborative filtering," In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999.

[5] Shardanand, U. and Maes, P., "Social Information Filtering: Algorithms for Automating "Word of Mouth"," In *Conference proceedings on Human factors in computing systems (ACM CHI '95)*, Denver, 1995, pp. 210-217.

[6] Hill, W., Stead, L., Rosenstein, M., and Furnas, G. W., "Recommending and Evaluating Choices in a Virtual Community of Use," In *Proceedings of ACM CHI'95 Conference on human factors in computing systems*, Denver, 1995, pp. 194-201.

[7] Dahlen, B. J., Konstan, J. A., Herlocker, J. L., Good, N., Borchers, A., and Riedl, J., "Jump-starting movielens: User benefits of starting a collaborative filtering system with "dead data"," University of Minnesota , TR 98-017, 1998.

[8] Goldberg, K., Roeder, T., Gupta, D., and Perkins, C.. Eigentaste: A Constant Time Collaborative Filtering Algorithm. Information Retrieval Journal . 2000.

[9] Schafer, J. B., Konstan, J. A., and Riedl, J., "Recommender systems in e-commerce," In *Proceedings of the ACM Conference on Electronic Commerce (EC-99)*, 1999, pp. 158-166.

[10] Morita, M. and Shinoda, Y., "Information filtering based on user behavior analysis and best match text retrieval," In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 272-281.

[11] Terveen, L., Hill, W., Amento, B., McDonald, D., and Creter, J., "PHOAKS: A System for Sharing Recommendations," *Communications of the ACM*, Vol. 40, No. 3, 1997, pp. 59-62.

[12] Breese, J., Heckerman, D., and Kadie, C., "Empirical analysis of predictive algorithms for collaborative filtering," In *Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence*, 1998.

[13] Pennock, D. M., Horvitz, E., Lawrence, S., and Giles, C. L., "Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- and Model-Based Approach," In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, UAI 2000*, Morgan Kaufmann, San Francisco, 2000, pp. 473-480.

[14] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R., "Indexing by latent semantic analysis," *Journal of the American Society of Information Science*, Vol. 41, No. 6, 1990, pp. 391-407.

[15] Ding, C. H. Q., "A similarity-based probability model for latent semantic indexing," In *Proceedings of the SIGIR*, ACM, 1999.

[16] Hofmann, T., "Probabilistic latent semantic indexing," In *Proceedings of the SIGIR*, ACM, 1999.

[17] Billsus, D. and Pazzani, M., "Learning Collaborative Information Filters," In *Machine Learning: Proceedings of the Fifteenth International Conference*, Morgan Kaufmann Publishers, San Francisco,CA, 1998, pp. 46-54.

[18] Pryor, M., "The effects of singular value decomposition on collaborative fitering," Dartmouth College CS, PCS-TR98-338, 1998.

[19] Lu, H., Lu, Z., and Li, Y., "EVS: Enhanced Vector Similarity for Collaborative Information recommendation based on SVD," In *International ICSC-NAISO Congress on Computaional Intelligence: Methods And Applications (CIMA 2001)*, 2001.

[20] Nichols, D. M., "Implicit Rating and Filtering," In *the Fifth DELOS Workshop on Filtering and Collaborative Filtering*, 1997, pp. 31-36.

[21] Oard, D. and Kim, J., "Implicit Feedback for Recommender Systems," In *Proceedings of the AAAI Workshop on Recommender Systems* , 1998.

[22] Claypool, M., Le, P., Waseda, M., and Brown, D., "Implicit Interest Indicators," In *Proceedings of ACM Intelligent User Interfaces Conference (IUI)*, Santa Fe, New Mexico, USA, 2001.

[23] Shapira, B., Shoval, P., and Hanani, U., "Experimentation with an information filtering system that combines cognitive and sociological filtering integrated with user stereotypes," *Decision Support System*, Vol. 27, 1999, pp. 5-24.

[24] Shapira, B., Shoval, P., and Hanani, U., "Stereotypes in Information Filtering Systems," *Information Processing and Management*, Vol. 33, No. 3, 1997, pp. 273-287.

[25] Basu, C., Hirsh, H., and Cohen, W., "Recommendation as classification: using social and content-based information in recommendation," In *Proceedings of the 1998 workshop on Recommender Systems*, AAAI Press, 1998, pp. 11-15.

[26] Cohen, W., "Fast Effective Rule Induction," In *Machine Learning: Proceedings of the Twelfth International Conference*, Morgan Kaufmann, California, 1995.

[27] Cohen, W., "Learning Trees and Rules with Set-valued Features," In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, Oregon, Portland, 1996.

[28] Joachims, T., Freitag, D., and Mitchell, T., "Webwatcher: A tour guide for the World Wide Web," In *Proceedings of IJCAI-97*, 1997.

[29] 路海明, 网络信息推荐的研究. [博士学位论文]. 北京：清华大学自动化系，2001

[30] Herlocker, J. L., Konstan, J. A., and Riedl, J., "Explaining Collaborative Filtering Recommendations," In *ACM 2000 Conference on Computer Supported Cooperative Work*, 2000.

Collaborative Filtering

Cai Deng，Lu Zengxiang，Li Yanda
Department of Automation, Tsinghua University,
Beijing 100084, China
caideng00@mails.tsinghua.edu.cn

Abstract：

With the rapid growth of the Internet, information overload is becoming a more and more serious problem. It is an urgent demand to help people to find useful information effectively. Of all the filtering technique, Automated collaborative filtering is quickly becoming a popular one for solving the problem. In this paper we try to uncover the essential of collaborative filtering, analyze some algorithm, and discuss the future of this technique.

Key words：information filtering, collaborative filtering, social filtering