VTT INFORMATION TECHNOLOGY

RESEARCH REPORT TTE4-2001-35

LOUHI-project

Collaborative Filtering and Recommendation Systems

Version 1.0

22.1.2002

Sonja Kangas

Version history

Version	Date	Author(s)	Reviewer	Description
0.1-1	13 Dec. 2001	Sonja Kangas		First draft
0.1-2	31 Dec. 2001	Sonja Kangas	Ville Ollikainen	
0.1-3	17 Jan. 2002	Sonja Kangas	Siv Pensar, Esa Rinta-Runsala	

Contact information

Sonja Kangas
VTT Information Technology
P.O. Box 12041, FIN-02044 VTT, Finland
Street Address: Tekniikantie 4 B, Espoo
Tel. +358 9 6052, fax +358 9 456 7024
Email: sonja.kangas@vtt.fi
Web: http://www.vtt.fi/tte/

UDK
Key Words Collaborative filtering, information filtering, targeted recommendations

Abstract

The purpose of this document is to introduce the wide area of intelligent filtering, mainly concentrating on collaborative filtering (CF), user profiling and recommendation systems. The background of collaborative filtering and its advantages and limitations, as well as examples and suggestions on how to improve collaborative filtering ideas for analyzing product data in the LOUHI-project have been presented in this report.

The development of collaborative filtering or its super-concept intelligent filtering is still ongoing. “Word-of-mouth” type of opinion and information sharing “systems” have been in use for ages, but from the first half of 1990s, the pervasion of the internet enabled new ways to carry out the idea of sharing opinion with a wide number of people via the net. The first known applications were Grouplens in 1992 and Firefly in 1994. Later Yahoo and Barnesandnoble signed up to use Firefly's technology. Finally book dealer Amazon.com introduced the idea of collaborative filtering (they had the BookMatcher system later spread to cover also other items) to a wider number of people in 1998. This is a basic knowhow report for the case studies within Phase 2 of the Louhi-project. Further development of the ideas will be implemented within the case studies in 2002-2003.

The major findings of this report are that collaborative filtering is a good possibility when building personalized recommendation systems. There are still many problems that have not been solved, e.g. scalability and reliability questions. Also the bi-directional development towards both user-based and item-based filtering have created new possibilities that have not yet been totally implemented. These are also the challenges for LOUHI cases. Collaborative filtering ideas also touch the ideas of semantic web. By semantics one can improve the abstraction level of recommendation systems. Besides comparing individual profiles one can also find other connections between the users and the items.[1]

[1] Viljanen, 2001.

Contents

Abstract 1

Contents 2

1 Introduction 3

1.1 Definitions 4

1.2 Recommendation systems 7

2 Filtering systems 9

2.1 Intelligent filtering and recommendation systems 10

Recommendation systems (RS) 10

2.2 User-based 10

2.2.1 Collaborative Filtering 11

2.2.2 Social Filtering 11

2.2.3 Clique-based filtering 11

2.2.4 Adaptive filtering 12

2.2.5 Others 12

2.3 Item-based 12

2.3.1 Feature-based filtering 13

2.3.2 Content-based filtering 13

2.3.3 Keyword-based filtering 13

2.3.4 Profile Filtering 14

2.4 Agents 14

3 Methods 15

3.1 Algorithms 15

3.2 Possibilities 21

3.3 Limitations 21

3.4 Privacy issues 22

3.5 Comparison of the methods 23

4 Conclusions 27

Bibliography 29

1 Introduction

This report is a start of a new perspective to LOUHI project objectives. During the year 2002 there will be done further development of collaborative filtering (CF) as well as CF case studies. This work is background information and a basic know-how report on recommendation systems and collaborative filtering.

Recommendation systems apply knowledge discovery techniques to the problem of making personalized recommendations for information, products and/or services. These systems, especially the ones based on nearest neighborhood collaborative filtering, have achieved widespread success on the web. First known applications were Grouplens in 1992 and Firefly in 1994. Later Yahoo and Barnesandnoble.com signed up to use Firefly's technology. Finally the book dealer Amazon.com introduced the idea of collaborative filtering (later their BookMatcher system spread to cover also other items) to wider audience in 1998.

The large growth in the amount of available information, the limitations of search engines like Altavista and Google, and the growing number of visitors to web sites in recent years pose some key challenges for recommendation systems. One should be able to find eligible data, which seems to be increasingly harder without some information filtering or recommendation systems. With recommendation systems the problems are e.g. how to produce high quality recommendations, perform many recommendations per second (real time) for largish amounts of users and items, and how to achieve high coverage in the face of data scarcity. In traditional collaborative filtering systems the amount of work increases with the number of participants and items in the system.

Work around different filtering ideas has been implemented widely, but very few general descriptions of the different systems are available. Most of the publications are presentations of different applications and thus all the information is seldomly given to the reader. Therefore one aim of this report is to give a somewhat general view on information filtering possibilities to day and also depict the future directions, what kind of potential collaborative filtering could offer to product data.

Malone et al.[2] describe three forms of information filtering: social (collaborative), content (cognitive) and economic. Social filtering has moved on from the original description (of the importance of the identity of the sender of a message) to several research projects and a few actively used systems. Currently social filtering is largely based on explicit ratings - where users rate a document on a pre-defined scale. Content-based filtering is dominant in information retrieval - typified by profiles based on keywords. Economic filtering is supposed to become increasingly important as digital cash; micro-payments and secure payment technologies emerge from research laboratories onto the internet.[3]

1.1 Definitions

Recommendation systems provide personalized suggestions about items that users might find interesting, by matching selections to user profiles or user groups. These systems require an interface that can "intelligently" determine the interest of a user and use this information to make suggestions. With collaborative filtering it is meant certain systems that “understand” its users. In other words we are trying to create systems wise enough to learn about their users, profile them and give suggestions. Users are individuals as well as members of a group (neighborhood). The grouping of items is a more complex issue as sometimes the problem of grouping items to product groups could possibly narrow the scale of recommendations. Below is a simplistic picture of recommendation systems.

```
Content                   Profile repository
(items)                   (users, ratings)
      |                   |
      |                   |
Content
search                    Profile matching,
engine                    creating neighborhoods
      \                   /
       \                 /
        \               /
Client interface (information source / user profiles)
```

Figure 1. Basic functions of recommendation systems

[2] Malone, Grant, Turbak , Brobst and Cohen, 1987.

[3] David M. Nichols, 1997.

a) Memory and model based CF

Recommendation system methods are divided into model based and memory-based algorithms[4]. Memory-based algorithms utilize the entire user-item database to generate predictions. Recommendation systems employ statistical techniques to find a set of users (neighbors) that have a history of agreeing with the target user (i.e., they either rate different items similarly or they tend to buy a similar set of items). Once a neighborhood of users is formed, recommendation systems use different algorithms to combine the preferences of neighbors to produce a prediction and/or top-N recommendation for the active user. The techniques, also known as nearest-neighbor or user-based collaborative filtering are more popular and widely used in practice than item-based systems.

Model-based collaborative filtering algorithms provide item recommendation by first developing a model of user ratings. Algorithms in this category take a probabilistic approach and envision the collaborative filtering process as computing the expected value of a user action, given his/her ratings on other items. The model building process is performed by different machine learning algorithms such as Bayesian network, clustering, and rule-based approaches.[5]

b) Explicit and implicit ratings

There are mainly two ways of collecting data and tracking the users' habits. First is explicit rating, where users *tell* the system what they think about a piece of information or certain products. Social filtering systems that use explicit ratings require a large number of ratings to remain viable. The effort involved for a user to rate a document may outweigh any benefit received, leading to a shortage of ratings. The idea of explicit filtering can be seen as a way to engage a user with the system. The difficulties to get a verification of the data users' give, is one disadvantage of this approach.

The other method is implicit rating. This way a rating is done by gathering information from *user behavior*. A server records user actions and concludes ratings from these actions. Implicit ratings include measures of interest such as

[4] Breese, Heckerman and Kadie, 1998.

[5] Sarwar, Karypis, Konstan, and Riedl, 2000.

click streams, whether the user has read an article and, if so, how much time the user spent reading it. The idea is to track users' behaviors. One issue to consider with implicit rating is, like it was already brought up at LOUHI workshop: how do we know when the user is buying/selecting things for her/himself and when she/he is buying e.g. a gift for someone.

In his Implicit Rating and Filtering[6] essay, David M. Nichols gives a list of several types of implicit data that can be captured and studied.

Action	**Notes**
Purchase (Price)	buys item
Assess	evaluates or recommends
Repeated Use (Number)	e.g. multiple check out stamps
Save / Print	saves document to personal storage
Delete	deletes an item in shopping basket
Refer	cites or otherwise refers to item
Reply (Time)	replies to item
Mark	add to a 'marked' or 'interesting' list
Examine / Read (Time)	looks at whole item
Consider (Time)	looks at abstract
Glimpse	sees title / surrogate in list
Associate	returns in search
Query	association of terms from queries

Figure 2. Potential types of implicit rating information

One alternative is to combine these two methods, users' actions (mouse clicks, mouse movements, scrolling and elapsed time, implicit ratings) are connected to user profiles done by the users' themselves (explicit ratings). This would also allow the system to probably give more specific random suggestions (the users can do some "window-shopping") on products that the users were not aware of or did not notice. In this way the system could give more extensive result on adapted suggestions. These kinds of systems are still unusual and there were not any useful empirical data available about these kinds of systems.

c) Passive and active systems

Collaborative filtering systems can be also divided to active and passive systems, on basis of the activity of the user requested. Some passive CF systems just ask the user to fill their user name before starting to offer suggestions and/or ask to make assessments about few items. Active systems can improve the use, ask background questions and form a group by that data or by the users action or both.

[6] David M. Nichols, 1997.

1.2 Recommendation systems

There are different types of recommendation systems based on the needs. Collaborative filtering and other filtering ideas belong to the recommendation systems category. A bit simpler systems are a) shared annotation systems, where a group of users can share notes e.g. concerning one particular document or a group of documents and b) information forwarding systems, where people can forward such information that could interest a certain group that has similar interests. ”Top N” systems are such where the user is being offered a list of the most popular documents in a given order. These recommendations do not base on any awareness of other users’ preferences but the overall popularity of certain documents. Often collaborative filtering systems use Top-N lists instead of giving customized or random recommendations.

The majority of collaborative filter based recommendation systems have been built by forming neighboring groups of people that seem to have somewhat similar taste or user/shopping needs/habits. This is carried through usually by forming:

1. predictions (based on users own actions and user history)
2. recommendation (own profile versus other users’ profiles)

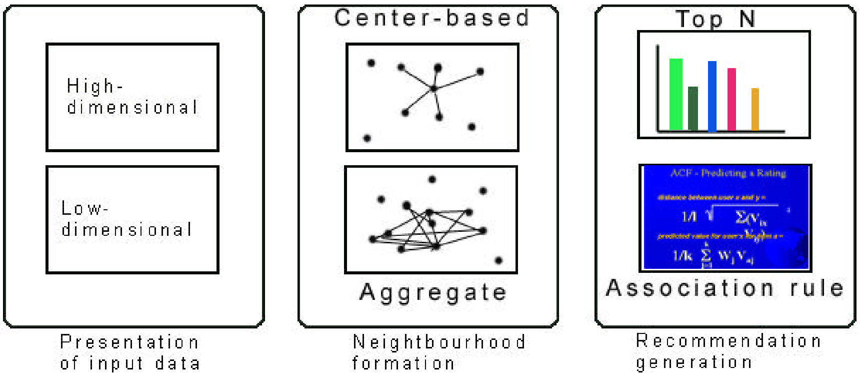


Figure 3. Basic structure of recommendation generation systems

There are several possibilities on how to create recommendations. The above picture is one kind of a view of the functionalities of recommendation making. There the neighborhood is either formatted by center-based: a schema forms a neighborhood for a particular customer by simply selecting the nearest other customers, or by aggregating: the schema forms a neighborhood for a customer by first picking the closest neighbors to that particular customer. Then the algorithm computes the rest of the neighbors, by calculating the cetroid of the neighborhood. Basically the algorithm allows the nearest neighbors to affect the formation of the neighborhood and it can be beneficial for very sparse data sets.[7] The recommendation is then generated either using Top N, association rule or combining them together by a certain rule, chosen by the developer of the system.

[7] The picture is modified by the example at:
http://www.cs.umn.edu/Research/GroupLens/ec00.pdf.

2 Filtering systems

There are various ways to perform personalized information filtering. Different machine learning algorithms can be used to learn a mapping from the features of an item to a number indicating the utility of the item to the user based on previous ratings that the user has made on other items. For example, the words in an article can form its features that can be used to predict whether an article could be interesting for the user.

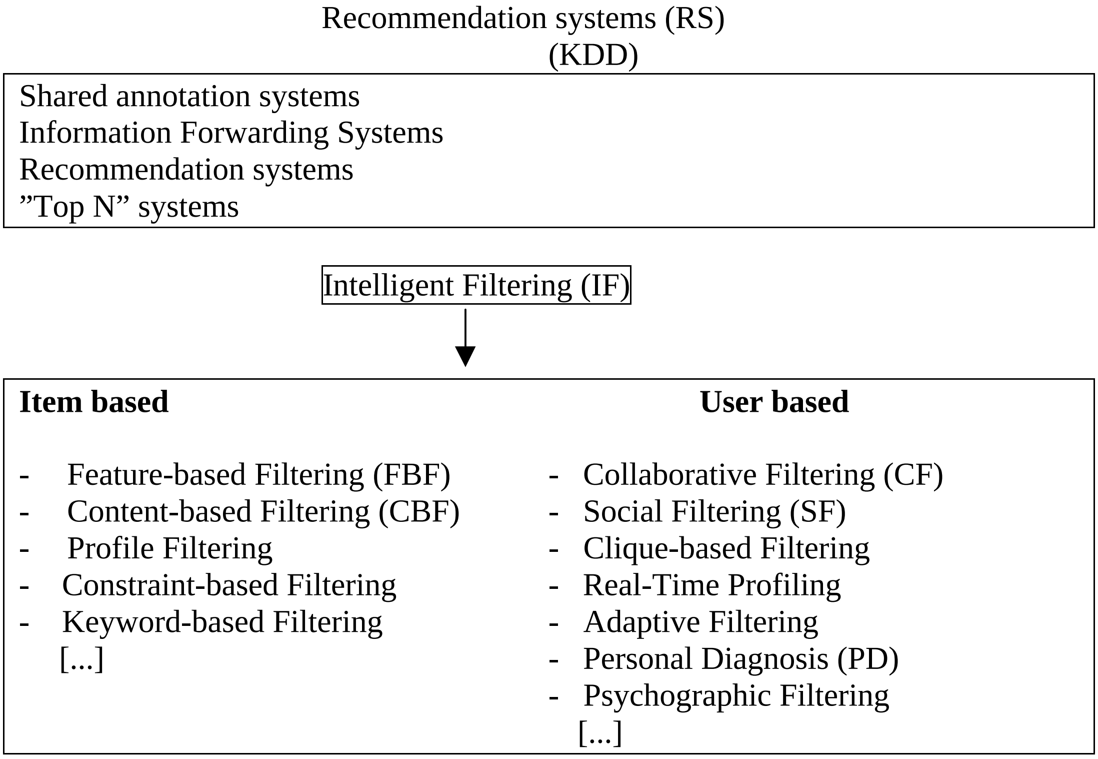
An alternative method is to use ratings of "similar" users in order to predict the rating on items that the user has not rated. This is called collaborative filtering. The basic premise is that people with similar taste tend to like similar types of items and the rating of someone similar is a good predictor for the personalized rating of the item.

Collaborative filtering can improve in taking advantage of the information. CF intensifies the user personalization from a search point of view, improves the accuracy of searches and decreases time that the user need to use for e.g. following up changed on a particular web site. The common examples of CF are systems that suggest books, articles, CDs or videos. One important thing from a product data point of view is that often information on books or CDs is valid for many years or even decades (evergreen, superstars, bestsellers). The database of items is extensive and actively used compared to material that changes or goes out of date rather quickly, such as news material, where the topics and headlines change many times even within a day. How to create worthwhile recommendations within such material?

The basic mechanism behind collaborative filtering systems is the following: a large group of people's preferences are registered using a similarity metric. Soon a neighborhood of people is selected, whose preferences are similar to the preferences of the person who seeks advice. Then a (possibly stressed) average of the preferences for that neighborhood is calculated and the resulting preference function is used to recommend options on which the advice-seeker has not yet expressed any personal opinion.[8]

[8] Heylighen, 1999.

2.1 Intelligent filtering and recommendation systems



Recommendation systems (RS)
(KDD)

Shared annotation systems
Information Forwarding Systems
Recommendation systems
”Top N” systems

Intelligent Filtering (IF)

Item based

- Feature-based Filtering (FBF)
- Content-based Filtering (CBF)
- Profile Filtering
- Constraint-based Filtering
- Keyword-based Filtering

[...]

User based

- Collaborative Filtering (CF)
- Social Filtering (SF)
- Clique-based Filtering
- Real-Time Profiling
- Adaptive Filtering
- Personal Diagnosis (PD)
- Psychographic Filtering

[...]

2.2 User-based

User-based CF has been the most frequently used technology for building recommendation systems to date, and is used in many commercial recommendation systems. The computational complexity of these methods increases linearly with the number of customers that sometimes in commercial applications can grow to large amounts as the systems are in the net and possibly the customers can be around the globe. Naturally national or local applications in Finland scarcely face this kind of problems. Scalability issues have been discovered to be one of the problem of user-based filtering. In able to avoid the scalability problems, recommendation making have been looked at the viewpoint of the items (item based CF). First different user-based collaborative filtering orientations will be described shortly.

2.2.1 Collaborative Filtering

Collaborative filtering compares your likes and dislikes to those of other people to predict your preferences. Based on the subjective evaluations of other readers, CF is a promising form of social filtering. A moderated newsgroup or virtual community, like duuni.net[9], employs a primitive form of collaborative filtering, choosing articles for all potential readers based on evaluations by a single person, the moderator. The moderator acts as “a filter” in these kinds of systems where the information is shared among peers to aid each other in finding the most interesting information.[10] One example of a collaborative filtering system is Amazon.com where large amounts of registered users review books. The system also gives suggestions on basis of one’s profile.

2.2.2 Social Filtering

Another form or often only a synonym for collaborative filtering is social filtering. In the basic level social filtering often means the same as collaborative filtering. Anyhow the terminology differs between various research groups. Social filtering is often compared to item-based CF systems. Social filtering overcomes some of the limitations of content-based filtering. A computer does not necessarily have to nderstand item properties. Furthermore, the system may recommend items that are very different (content-wise) from what the user has indicated liking before. Finally, recommendations are based on the quality of items, rather than more objective properties of the items themselves[11].

2.2.3 Clique-based filtering

Clique-based filtering is yet another term for collaborative filtering and often it is used as a synonym for collaborative filtering. The clique-based filtering approach uses a group of similar-minded people as indicators of a user's interests. The assumption is that users who feel similarly about previous items will feel similarly about new items.

[9] http://duuni.talentum.com/.

[10] Berst, 1997.

[11] Berst, 1997.

2.2.4 Adaptive filtering

Adaptive filtering is a kind of a combination of user-based and item-based filtering. There the idea is that the system learns as it goes along, by asking the user to rate things and by monitoring the click stream to watch what the user does. For instance, the search service Excite has a News Tracker service that asks you to check the stories you liked and then hit a "learn" button to fine tune your preferences. Wisewire.com uses a similar method, combining it with collaborative filtering as well.[12]

2.2.5 Others

There are several different terms for user-based filtering “genres”. As examples here is mentioned psychographic filtering and personality diagnosis. Psychographic filtering is similar to collaborative filtering, except that it predicts your likes and dislikes based on a "psychographic profile" derived from a questionnaire. The Affinicast Interaction Manager is a leading example of this approach. In personality diagnosis (PD) is a way, when given a user's preferences for some items, one computes the probability that he or she is of the same "personality type" as other uses, and, in turn, the probability that he or she will like new items.

2.3 Item-based

The other perspective to filtering is item based. Such item-based CF recommendation techniques have been developed that analyze the user-item matrix to identify relations between the different items, and use these relations to compute the list of recommendations. First the similarities or differences between items are searched, then those evaluations are connected to users or a group of users.

The bottleneck in user-based collaborative filtering algorithms is the search for neighbors among a large user population of potential neighbours.[13]

[12] Berst, 1997.

[13] Sarwar, 2001.

Recommendations for users are computed by finding items that are similar to other items the user has liked. Because the relationships between items are relatively static, item-based algorithms may be able to provide the same quality as the user-based algorithms with less online computation. E.g. the joke recommendation system Jester is an example of an item-based recommendation system.

2.3.1 Feature-based filtering

Feature-based approach is based on the idea that it is possible to capture the features a user likes and does not like about an item and thus provide feedback about various items to the user. Feature-based filtering is also used when talking about e.g. movies. Movies could be filtered on basis of the director, actor, theme, and happy/sad end. The filtering system would understand the features of an item, in this case movie, and make recommendations on basis of user profile or retrieval of certain types of movies. The limitation is that often the features can only be retrieved from text data.

2.3.2 Content-based filtering

In content-based filtering (also called cognitive filtering), document representations can exploit only the piece of information that can be derived from document contents. We need technology to help us wade through all the information to find the items we really want and need, and to rid us of the things we do not want to be bothered with. Content-based filtering techniques recommend items for the user's consumption based on correlations between the content of the items and the user's preferences. For example, the system may try to correlate the presence of keywords in an article with the user's taste.

2.3.3 Keyword-based filtering

Keyword-based filtering is a simple version of content-based filtering. It is limited to the kinds of concepts that can be expressed in terms of keywords. As a result, there may be some news articles for example, which will be difficult to retrieve using simple keyword-based searches.

2.3.4 Profile Filtering

Profile filtering is the most straightforward approach. The users describe their interests by picking from a list or entering keywords, and the software rejects anything that doesn't match. ZDNet's Personal News Service uses this approach. Many other news sites have similar features.[14]

2.4 Agents

Intelligent agent technology is one alternative for recommendation systems. The development on this area is mainly done at Massachusetts Institute of Technology (MIT) in USA. As an agent they mean an autonomous software process which acts on behalf of a client. They also use a term “social agents”, when meaning social relationship between the clients. They have a strong belief that next generation intelligent information systems will rely on cooperative agents for playing a fundamental role in actively searching and finding relevant information on behalf of their users in complex and open environments, such as the internet.

E.g. the collaborative filter system FAB uses agents. The idea of this recommendation system is to recommend web pages for certain users. It profiles the users as well as web pages using collector and chooser agents.

[14] Berst, 1997.

3 Methods

Within this section it is presented what are the qualities of CF, how it is implemented and also some examples are given of what kinds of CF systems there are. There is also the comparison of the methods used. At this point we are not developing any new ideas, just reviewing and comparing the available applications (passive/active, implicit/explicit, user/item-based, filtering).

3.1 Algorithms

CF algorithms should be able to give personalized suggestions by using evaluations from the users that have the same preferences. Various algorithms have been used with collaborative filtering, both recommendation systems and similarity measures. Often Top N ratings (average of all users) are differentiated from CF systems, even though Top N systems are used in a smaller scale by concentrating on certain user’s Top N selections prorata neighborhoods. Other technologies have also been applied to recommendation systems, including Bayesian networks, Pearson correlation, cosine correlation, clustering and horting.

a) Clustering

Clustering techniques usually produce less personal recommendations than other methods, and in some cases, the clusters have worse accuracy than nearest neighbor algorithms. Once the clustering is complete, however, the performance can be very good, since the number of the groups that must be analysed is smaller. Clustering techniques can also be applied as a ”first step” for shrinking the candidate set in a nearest neighbor algorithm or for distributing nearest-neighbor computation across several recommendation engines. While dividing the population into clusters may hurt the accuracy of recommendations for users near the fringes of their assigned cluster, pre-clustering may be a worthwhile trade-off between accuracy and efficiency[15].

[15] Sarwan, 2001.

b) Horting

Horting is a graph-based technique in which nodes are users, and edges between nodes indicate degree of similarity between two users. Predictions are produced by walking the graph to nearby nodes and combining the opinions of the nearby users.[16]

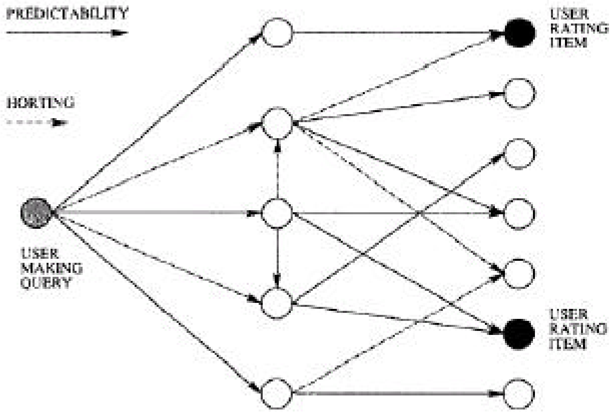


Figure 4: The nodes are the users and the edges predictability. Horting means the similarities between different users.

Horting differs from nearest neighbor algorithms as the graph may be walked through other users who have not rated the item in question, thus exploring transitive relationships that nearest neighbor algorithms do not consider. In one study using synthetic data, horting produced better predictions than a nearest neighbor algorithm.

Horting has been mentioned to be appropriate e.g. for e-commerce merchants offering one or more neighborhoods of relatively homogenous items. Schafer et al. (1999), present a taxonomy and examples of recommendation systems used in e-commerce and how they can provide one-to-one personalization and capture customer loyalty. Although these systems have been successful in the past, their widespread use has exposed some limitations such as problems of sparsity in the data set and problems associated with high dimensionality.[17]

[16] Aggarwal, Wolf, Wu and Yu, 1999. Figure 4 from the same source.

[17] Sarwan, 2001.

c) Bayesian belief network

Bayesian Belief Networks (BBN) are also known as Belief Networks, Causal Probabilistic Networks, Causal Nets, Graphical Probability Networks, Probabilistic Cause-Effect Models, and Probabilistic Influence Diagrams. Bayesian logic is a branch of logic applied to decision making and inferential statistics that deals with probability inference: using the knowledge of prior events to predict future events. To date BBNs have proven useful in practical applications such as medical diagnosis and diagnosis of mechanical failures.

A BBN is a graphical network that represents probabilistic relationships among variables. BBNs enable reasoning under uncertainty and combine the advantages of an intuitive visual representation with a sound mathematical basis in Bayesian probability.

$p(A|B) = p(A, B)/p(B)$

Bayesian networks create a model, which is based on user information. It may be feasible for environments in which the knowledge of user preferences is changing slowly with respect to the time needed to build the model but these models are not suitable for environments in which user preference models must be updated rapidly or frequently.

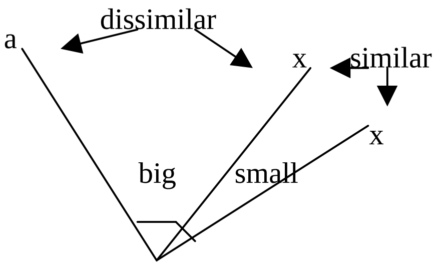
d) Genetic (evolutionary) algorithm

Genetic Algorithms (GA) are methods of function maximization which mimic the selective and mutative properties of the natural selection among biological organisms. They are especially well suited to problems which can not be solved by mathematical evaluation and for which direct search in the solution space is prohibitively expensive.[18] In the case of collaborative filtering, the idea of a genetic algorithm is that the computer presents choices (species in GA). Then the user selects the preferred ones, and the computer iterates using that further information - in other words, design by multiple selection. Collaborative technology uses other people's inputs to help refine the system's selections at each stage[19].

[18] Wall, 1999.

[19] Colwell, 1996.

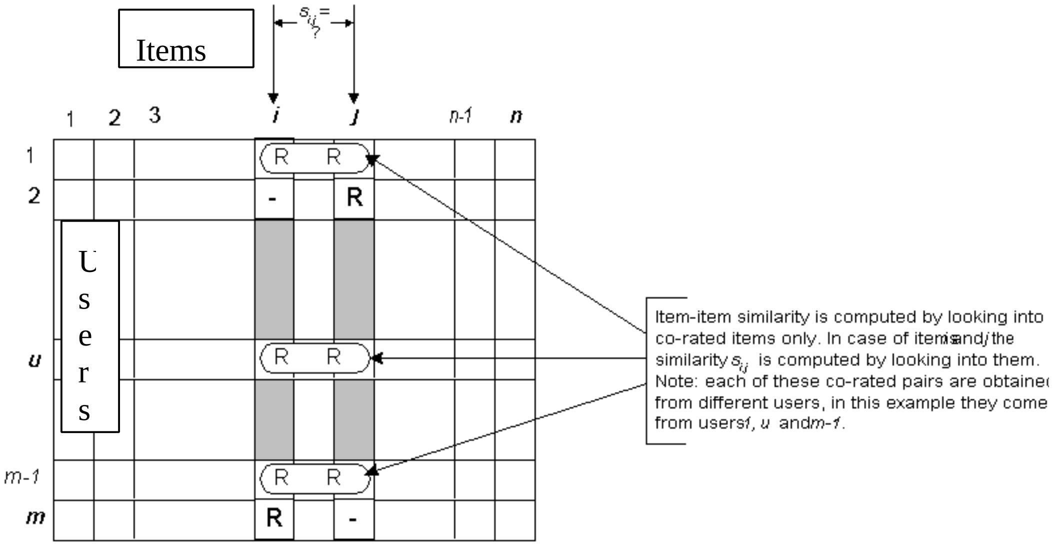
e) Cosine-based Similarity



$$sim(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|_2 * \|\vec{j}\|_2}$$

Figure 5. Cosine algorithm

The cosine algorithm gives a good similarity measure for two vectors in a multi-dimensional space. The space may describe users or characteristics, such as keywords, of items and the items are thought of as vectors in this space. The similarity between items is measured by computing the cosine of the angle between these two vectors[20]. The cosine algorithm provides a similarity measure between 0 and 1.



Item-item similarity is computed by looking into co-rated items only. In case of items i and j the similarity s_{ij} is computed by looking into them. Note: each of these co-rated pairs are obtaine from different users, in this example they come from users 1, u and m-1.

Figure 6: An example of cosine-based similarity.

f) Pearson correlation

One typical similarity metric is the Pearson correlation coefficient between the users' preference functions or (less frequently) vector distances or dot products. There are several formulas for Pearson's correlation. A commonly used formula is shown below. The method is used for measuring the linear similarity between two user profiles or items etc. by using vector representation, to determine the angle between the vectors. The formula gives an approximation on how well the vectors

[20] Sarwar, 2000. Also images are from the same source, Goffinet&Noirhomme-Fraiture, 1995.

(ie profiles, items) compared, match from the scale of zero (no similarity) to one (match totally) or minus one (totally different).

$$r = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{(\Sigma X^2 - \frac{(\Sigma X)^2}{N}) \quad (\Sigma Y^2 - \frac{(\Sigma Y)^2}{N})}}$$

Figure 4. Pearson correlation

g) Weighted Sum

The most important step in the collaborative filtering systems is to generate the output interface in terms of prediction. Once the set of the most similar items based on the similarity measures is isolated, the next step is to look into the target users ratings and use a technique to obtain predictions.

$$P_{u,i} = \frac{\sum_{\text{all similar items, N}} (s_{i,N} * R_{u,N})}{\sum_{\text{all similar items, N}} (|s_{i,N}|)}$$

Figure 8: Weighted Sum

Basically, this approach tries to capture how the active user rates the similar items. The predicted rating for an item is got by weighting each of the items rated by its similarity to the target item and summing the results. The weighted sum is scaled by the sum of the similarity terms to make sure the prediction is within the predefined range.[21] Also other algorithms have been used. These are just a few examples of user- and item-based filtering algorithms to give a picture of the functionality of different algorithms.

h) Regression

The regression approach is near the weighted sum technique. The main difference is that the raw ratings of items are replaced by their approximate values based on a linear regression model. This is because in practice, the similarities computed

[21] Sarwar, 2001.

using cosine or correlation measures may be misleading. The ratings vectors may be distant in Euclidean sense but they may end up with very high similarity value. A prediction based on these occasionally misleading similarity measures may be poor and regression technique is intended to correct this.[22]

[22] Sarwar, 2001.

3.2 Possibilities

The possibilities of CF imply:

- more loyal customers
- growth of sales
- constant and detailed feedback of the products from the users
- better understanding of the markets
- products to costricted clientele
- directed advertising
- saves money when easy to find effectively needs/wants

It has been taken into account that collaborative filtering works well for a closed set of items (e.g. music, books). It will not work as well for online services seeking to recommend relevant news stories or articles for purchase because the number of possible "subjects" is too large and diverse. It is also difficult to infer value from user rankings beyond the immediate set of items being ranked. (Does indicating a preference correlate with drinking Pepsi rather than Coke?) The technology cannot take personal idiosyncrasies into account. The effectiveness of the collaborative filtering system depends also on the size of the sample, capture of data, speed and algorithms.

With collaborative filtering there has been done a rather good result on recommending items, but it takes some time before there are, in the database, enough user profiles and items to recommend. Another related question is, how to handle novelties, compared to evergreen or items dusting in the archives. Available user experiences are mainly built up by system developers and thus may be biased.

3.3 Limitations

Currently there can also be named several limitations on using CF:

- suitable graduation scale
- motivation of evaluators
- partiality of evaluators
- incentive (why should anyone rate anything?)
- avoid ”free-riding” problem
- reach the critical mass of evaluators

One problem with a recommendation system is that the system can produce a valuable recommendation only after it has accumulated a large set of ratings (but the likelihood of shared documents/items between different users could decrease). Also the general problems of information overload and low signal to noise ratio have received considerable attention in the research literature.

In practice, many commercial recommendation systems are used to evaluate large item sets (e.g. CDnow.com recommends music, MovieCritic movies, videos). In these systems, even active users may have purchased well less than 1% of the items. Accordingly, a recommendation system based on nearest neighbour algorithms may be unable to make any item recommendations for a particular user. Nearest neighbour algorithms require computation that grows with both the number of users and the number of items. With large number of users and items, a typical web-based recommendation system running existing algorithms will suffer serious scalability problems.

Within item-based CF systems, one problem is that the content can only be in machine parsable form, often meaning text. Sound, photographs or video, whether containing important on content or not, cannot be used. Content-based filtering methods cannot filter items based on quality, style or point-of-view. For example, they cannot distinguish between a well written and a badly written article if the two articles happen to use the same terms.

3.4 Privacy issues

In able to get the highest benefit from collaborative filtering, the system should tell where people go on the web, where they click, how long their session lasts, how they rate things, what kind of persons they are etc. This is what the system developers want to know in order to create better recommendations. The customers want more personalized suggestions too, but want their privacy at the same time. Certain questions about privacy and collaborative filtering have already been solved by the development of more general web standards.

One solution is Open Profiling Standard (OPS). It was brought to the World Wide Web Consortium (W3C[23]) as a way to give the control of data back to the user. Collaborative filtering requires a large amount of data to facilitate the correlation of a consumer's purchasing pattern, and recommendations cannot be changed or

[23] http://www.w3c.org.

weighted more heavily by an expert's opinion. Thus, collaborative filtering is appropriate when decisions are not of significant importance, and is used primarily for products that are influenced by word-of-mouth. [24] In this way, web sites would use personalized software like collaborative filtering, while users could protect their privacy. The technology behind OPS was a combination of the marketing powers of collaborative filtering and the application of the W3C's RDF (Resource Definition Framework).

3.5 Comparison of the methods

The scale of CF application is rather wide. Most popular CF systems are book, movie, news, CD etc. filtering systems: items that stay "current" for a relatively long time and belong to distinctly defined groups. Naturally also other types of items and filtering systems can be found. Subsequently a few CF applications have been grouped by the type of items because there are lots of similarities between different applications. Thus on the comparison table, we will only look at from the categorical point of view. Categories are made on basis of the type of the applications and filtered material. Mathematical methods might be different but at this point that does not matter that much as we are just looking at the principle working patterns.

Categories:

1. Movie: Firefly, MovieCritic[25], Täsmäase, MovieLens[26], Mangarate[27], Morse
2. Music: Firefly, CdNow[28]
3. Books: Amazon[29], Barnes&Noble[30]
4. Web: Webwatcher[31], Webfilter, Webwasher, Select[32], Webdoggie, Gustos[33]
5. Library/museum: ScienceIndex[34], Active Web Museum[35], BIRD, ChaffAway

[24] Look e.g.: http://www.sims.berkeley.edu/resources/collab/collab-report.html.

[25] http://www.finnkino.fi/rekisteroityminen/mc-esittely.asp.

[26] http://movielens.umn.edu/.

[27] http://hal.ddns.comp.nus.edu.sg/mangarate/htmldocs/.

[28] http://www.cdnow.com/.

[29] http://www.amazon.com/exec/obidos/subst/home/home.html/002-8486164-1136030.

[30] http://www.barnesnoble.com/.

[31] http://www.internetwatcher.com/Eng/Index.htm.

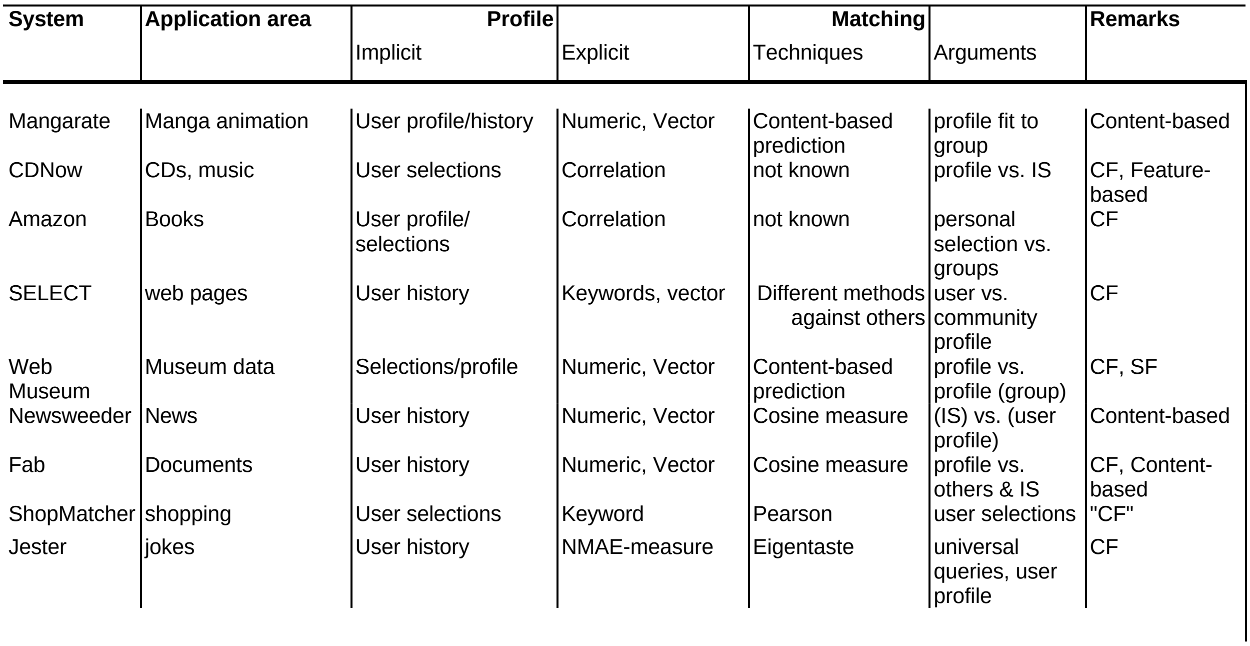
[32] http://www.dsv.su.se/~jpalme/select/select.html.

[33] http://www.gustos.com/.

6. News: Shift, Infoscan, NewsSieve, Borger, RAMA, Grouplens[36]
7. Documents: Fab
8. E-commerce: TripMatcher (travel), ShopMatcher (shopping), E-Markets
9. Other: Restaurant recommendation system (WAP)[37], Footprints, Jester (jokes)[38], JobMatcher (employment), WeddingNetwork (weddings), Levis (jeans), Yenta (matchmaking)[39], Trabble (restaurants).

Comparison topics:

- application area
- profile (implicit/explicit)
- mathematical method (matching techniques, arguments)
- make good use of product data
- real-time
- quality standard of data



System	Application area	Profile		Matching		Remarks
		Implicit	Explicit	Techniques	Arguments	
Mangarate	Manga animation	User profile/history	Numeric, Vector	Content-based prediction	profile fit to group	Content-based
CDNow	CDs, music	User selections	Correlation	not known	profile vs. IS	CF, Feature-based
Amazon	Books	User profile/ selections	Correlation	not known	personal selection vs. groups	CF
SELECT	web pages	User history	Keywords, vector	Different methods against others	user vs. community profile	CF
Web Museum	Museum data	Selections/profile	Numeric, Vector	Content-based prediction	profile vs. profile (group)	CF, SF
Newsweeder	News	User history	Numeric, Vector	Cosine measure	(IS) vs. (user profile)	Content-based
Fab	Documents	User history	Numeric, Vector	Cosine measure	profile vs. others & IS	CF, Content-based
ShopMatcher	shopping	User selections	Keyword	Pearson	user selections	"CF"
Jester	jokes	User history	NMAE-measure	Eigentaste	universal queries, user profile	CF

Figure 9: Comparison table of different filtering systems (by categories)

[34] http://www.scienceindex.com/.

[35] http://www.eurecom.fr/~kohrs/avwm/.

[36] http://www.cs.umn.edu/Research/GroupLens/.

[37] http://lcs.www.media.mit.edu/groups/agents/projects/.

[38] http://shadow.ieor.berkeley.edu/humor/.

[39] http://foner.www.media.mit.edu/people/foner/Yenta/.

The examples given were mainly for items that were not tightly tied to a certain time scale, but also the items compared were rather diverse. Manga animations act as a bounded genre of comic strip. The users of the system are somewhat known as only those who know about Japanese Manga comics/movies, has some use of the filtering system, unless the user wants to learn about Manga and just blindfolded starts looking for recommendations. The other end is Amazon and CDNow, where the variety of items is so wide that there is “something for everybody”. Books and news material is the other end of a line. Books stay “valid” for a longer period of time than the news. Also old vs. new data handling is carried out differently partly because of the circulation of the data/items is so different.

These systems lean mainly on user profiles or selections (keywords, user selections). Some systems expect the user to rate things sufficiently to be able to generate good enough recommendations. The problem with this kind of systems (e.g. MovieCritic and Jester) is that few users have the will or ability to rate enough items. The motivation to rate things is like to decrease after a few dozens of ratings and the system still does not understand your taste. Rating takes quite a lot of time and often the ratings would be better if the user would continually visit the website and rate new things.

As it has been mentioned earlier, there is not enough reliable information available on user experiences, so here are some general impressions of the user experiences. The amount of profile data given to the systems was often very limited. The privacy issue was not that big of a question as the idea was to get personalized recommendations. In able to get personalized recommendations, the users have to give some information in able to get good enough recommendations. Currently available collaborative filtering systems works diversively. Most of them learn by the amount of ratings done by the user. Others are mainly based on Top N type of classifications. The system gives recommendations by certain selections from pre-defined alternatives.

Both systems have their positive and negative sides. The first-mentioned systems assume that users will recurrently rate new items. This might be the case if the user really is interested about the items. An average user probably will not bother to review enough items in which case also the recommendations become inadequate. Item-based filtering would be one solution for this matter, that has not been taking in use in current CF systems. Also other solutions are available. Top N systems work well if one has clear enough alternatives where to choose. Such systems have been used for example for choosing presents or toys.

Again one should point out that these comments concern current systems where all the possible alternatives have not been taken into considerations. The general estimation is that when implemented well enough, collaborative filtering gives interesting and valuable possibilities to customer/user personalization and would be valuable to both customers and makers.

4 Conclusions

Collaborative filtering is a rather new technology used in the internet and is meant is be a “ light” alternative to e.g. data mining (e.g. neural networks, rule-based systems). It is full of possibilities that are not yet fully exploited, also lots of things are still under development or testing. On the internet the semantic web[40] approach may give interesting viewpoints to collaborative filtering and recommendation making. Also it has been discussed whether digital television could be a good ground for collaborative filtering. For digital television[41] collaborative filtering could be used for e.g. recommending programs or movies. The picture systems would use implicit ratings in a way that they track the users habits of watching television and on basis of that would create recommendations on coming up programs or movies.

New terminals such as mobile devices and digital television will give new possibilities as well as alternatives to collaborative filtering. On the net it has already achieved trust and high enough amount of users for some systems. Future development of the collaborative filtering scheme will draw wider its possibilities.

Despite of these encouraging words, there are still lots to be done from the research, methodological as well as product development point of views. There are clear benefits of using collaborative filtering for certain systems, but currently it is not applicable for every kind of items. The major problems are: how to build collaborative filter systems for material that changes often or is diverse, how to understand little differences (for example differences between religions (it is not advisable to suggest Christian literature to Muslim) or different cola drinks (Coca Cola fans do not want to get suggestions about any other Cola brand).

Privacy has been one of the central topics when talking about the problems with collaborative filtering. Partly the problem has already been solved as general web standards have been originated. Still the privacy will be one main area of discussion when developing collaborative filtering systems.

[40] Semantic web expresses information in a abstracted model with serializable properties. Semantic web elements have URI and it expresses information about web services. Semantic web models will accumulate more information as time progresses.

[41] See eg.: ReplayTV, http://www.replay.com/ and TiVo http://www.tivo.com/home_flash.asp.

Intelligent filtering possibilities give a new perspective to the LOUHI project questions and a new way to look at personalization and targeting items or offers to a particular customer or a group of customers. The cases carried out in 2002 will eventually show the actual benefits of recommendation systems for the LOUHI project.

Bibliography

Aarnio, Elias (2001). *Microsoftin digi-tv kyttää katsojaa*. Digitoday, 12.12.2001. URL: http://www.digitoday.fi/digi98fi.nsf/pub/md20011212145053_ea_6583998.

Aggarwal, C. C., Wolf, J. L., Wu K., and Yu, P. S. (1999). Horting Hatches an Egg: A New Graph-theoretic Approach to Collaborative Filtering. In *Proceedings of the ACM KDD'99 Conference*. San Diego, CA. pp. 201-212.

Alspector, J, A. Kolcz, and N. Karunanithi (1997). *"Feature-based and Clique-based Models for Movie Selection: A Comparative Study,"* User Modeling and User Adapted Interaction, vol. 37, no. 4, pp. 279-304.

Berst, Jesse (1997). *Intelligent Filters To the Rescue! (Sort of...)* ZDNet AnchorDesk. URL: http://www5.zdnet.com/anchordesk/story/story_1035.html.

Boersma, Peter, Boeve, Eddy and Wessels, Charles (2000): *Branding, Trust and User Experience in 1-to-1 E-commerce.* Amsterdam, The Netherlands. Position paper for the CHI2000 workshop "Designing Interactive Systems for 1-to-1 E-commerce".

Breese, Heckerman and Kadie (1998). *Empirical Analysis of Predictive Algorithms for Collaborative Filtering* in Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI, July. Morgan Kaufmann Publisher, pp. 43-52.

Charlet, Jean-Charlet (1998). *FireFly Network.* Stanford University Graduate School of Business Case Study.

Charlet, Jean-Charlet (1998). *Collaborative Filtering*. Stanford University Graduate School of Business Technology Notes.

Colvell, Steve (1996). *Collaborative filtering*. URL: http://stevecolwell.com/collfilt.html.

Gardner, Elizabeth (1999). *Sites Invest Big in Technology, People Improve Service*. Internet World.

ERCIM – Fifth DELOS workshop (1997), *Filtering and Collaborative Filtering*. Budapest 10-12, November 1997. URL: http://www.ercim.org/publication/ws-proceedings/DELOS5/delos5.pdf /.

Foltz, P.W. and Dumais, S.T. (1992). *Personalized information delivery: an analysis of information filtering methods,* Communications of the ACM, 35(12), pp. 51-60.

Goffinet, Luc & Noirhomme-Fraiture, Monique: *Automatic Hypertext Link Generation based on Similarity Measures between Documents*. Institut d'Informatique, FUNDP, 1995. URL: http://www.fundp.ac.be/~lgoffine/Hypertext/semantic_links.html.

Heylighen (1999). *Collaborative Filtering* in F. Heylighen, C. Joslyn and V. Turchin (editors): Principia Cybernetica Web (Principia Cybernetica, Brussels), URL: http://pespmc1.vub.ac.be/COLLFILT.html.

Hoffman, Donna L., Thomas P. Novak (1996). *Marketing in Hypermedia Computer-Mediated Environments: Conceptual Foundations*. Journal of Marketing. (July 1996) Vol. 60, pp. 50-68.

Honkala, Marko (1997). *Henkilökohtaistaminen selailukäyttäytymisen pohjalta*. Diplomityö (dissertation) HUT.

Kohrs, Arnd and Merialdo, Bernard (1999): *Improving Collaborative Filtering with Multimedia Indexing Techniques to create User-Adapting Web Sites*. Institut EURECOM, Department of Multimedia Communications. URL: http://www.kom.e-technik.tu-darmstadt.de/acmmm99/ep/kohrs/.

Malone, T.W., Grant, K.R., Turbak, F.A., Brobst, S.A. and Cohen, M.D. (1987). *Intelligent information sharing systems*, Communications of the ACM, 30(5), pp-390-402.

Morita, M. and Shinoda, Y. (1994). *Information filtering based on user behaviour analysis and best match text retrieval.* Proceedings of the 17th ACM Annual International Conference on Research and Development in Information Retrieval (SIGIR'94), Dublin, Ireland, Springer-Verlag, pp. 272-81.

Neil, Martin (2000). *Bayesian Belief Networks*. URL: http://www.agena.co.uk/bbn_article/bbns.html.

Nichols, David M (1997). *Implicit Rating and Filtering*. Computing Department, Lancaster University. In Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering, Budapest, Hungary, 10-12 November 1997, ERCIM, pp. 31-36.

Oard, D.W. and Marchionini, G. (1996). *A Conceptual Framework for Text Filtering*, Technical Report CAR-TR-830, Human Computer Interaction Laboratory, University of Maryland at College Park.

Palme, J. (1997). *Choices in the Implementation of Rating*, in Alton-Scheidl, R., Schumutzer, R., Sint, P.P. and Tscherteu, G. (Eds.), Voting, Rating, Annotation: Web4Groups and other projects: approaches and first experiences, Vienna, Austria: Oldenbourg, pp. 147-62.

Olson, Peter, J. Paul, Jerry C. (1999). *Consumer Behavior and Marketing Strategy*, Irwin McGraw-Hill.

Sarwar, Bardur (2001). *Collaborative Filtering Based Recommender Systems*. URL: http://www10.org/cdrom/papers/519/node1.html.

Sarwar, Karypis, Konstan, Riedl (2000). *Analysis of Recommendation Algorithms for E-Commerce*.

Sarwar, Karypis, Konstan, and Riedl (2001). *Item-based Collaborative Filtering Recommendation Algorithms.* Department of Computer Science and Engineering. University of Minnesota.

Schafer, J. B., Konstan, J., & Riedl, J. (1999). *Recommendation systems in e-commerce*. In Proceedings of the ACM Conference on Electronic Commerce, pp. 158-166.

Stevens, C. (1992). *Knowledge-Based Assistance for Accessing Large, Poorly Structured Information Spaces*, PhD Thesis, Department of Computer Science, University of Colorado.

Viljanen, Kim (2001). *Monilähteinen suosittelu,* Tietojenkäsittelytieteen laitos, Helsingin yliopisto (preliminary version of his master thesis).

Wall, Matthews (1999). *Introduction to Genetic Algorithms*. URL: http://lancet.mit.edu/~mbwall/presentations/IntroToGAs/index.html

Wittenburg, K., Das, D., Hill, W.C. and Stead, L. (1995). *Group asynchronous browsing on the World Wide Web*, Proceedings of the Fourth International World Wide Web Conference, Boston, MA, O'Reilly & Associates, pp. 51-62.