Auton Agent Multi-Agent Syst (2008) 17:251–269
DOI 10.1007/s10458-008-9029-x

User evaluation of a market-based recommender system

Yan Zheng Wei · Nicholas R. Jennings · Luc Moreau · Wendy Hall

Published online: 30 January 2008

Abstract Recommender systems have been developed for a wide variety of applications (ranging from books, to holidays, to web pages). These systems have used a number of different approaches, since no one technique is best for all users in all situations. Given this, we believe that to be effective, systems should incorporate a wide variety of such techniques and then some form of overarching framework should be put in place to coordinate them so that only the best recommendations (from whatever source) are presented to the user. To this end, in our previous work, we detailed a market-based approach in which various recommender agents competed with one another to present their recommendations to the user. We showed through theoretical analysis and empirical evaluation with simulated users that an appropriately designed marketplace should be able to provide effective coordination. Building on this, we now report on the development of this multi-agent system and its evaluation with *real users*. Specifically, we show that our system is capable of consistently giving high quality recommendations, that the best recommendations that could be put forward are actually put forward, and that the combination of recommenders performs better than any constituent recommender.

Keywords Recommender systems · Auctions · Marketplace · User evaluation

Y. Z. Wei
Department of Broadband Wireless Management, Huawei, B1-F2-B, Huadian, Bantian, Shenzhen 518219, China
e-mail: weiyanzheng@huawei.com

N. R. Jennings (✉) · L. Moreau · W. Hall
School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK
e-mail: nrj@ecs.soton.ac.uk

L. Moreau
e-mail: l.moreau@ecs.soton.ac.uk

W. Hall
e-mail: wh@ecs.soton.ac.uk

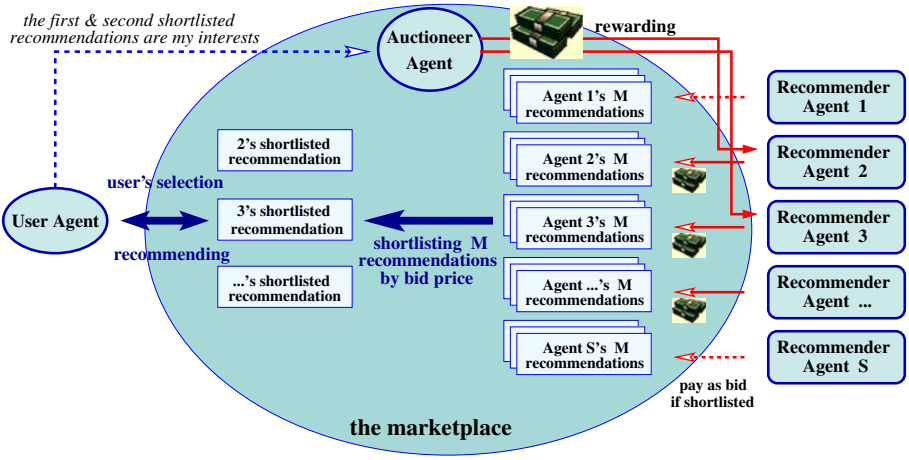

Fig. 1 The marketplace architecture

1 Introduction

Recommender systems have been widely advocated as a way of coping with the problem of information overload in many application domains and, to date, many recommendation methods have been developed. Such methods can broadly be classified into two main categories based on the attributes of the recommendations they consider [7]: (i) content-based filtering and (ii) collaborative filtering. The former work on the objective attributes of the recommendations (such as the textual contents of an article), whereas the latter work on the subjective ones (such as who else likes it) [4]. In either case, however, recommendations are made by the underlying method predicting the users' preferences for the various possible items that could be put forward. In addition, in many real world contexts, users may pay attention to either or both of the objective and the subjective attributes of the recommendation items. For example, when seeking an online movie, a user's attention may focus on either the objective textual introduction to the movie or other users' subjective ratings on it (or both). Therefore, different recommendation methods are likely to perform with varying effectiveness for different users in different situations. In short, there is no universally best method.

To combat this, we believe the way forward is to have a pool of constituent recommenders (each based on a particular method) and then provide an overarching framework that coordinates them so that only the best recommendations (from whatever source) are presented to the user. To this end, we have previously specified a system that recommends relevant online documents (represented as URLs) in which this coordination is achieved via a marketplace (see Fig. 1) in which recommender agents compete with one another to have their suggestions placed before the user. See [13] for a detailed justification of the choice of a market-based approach for this problem and for a detailed comparison with the state of the art in recommender and multi-agent systems.

In more detail, when a user visits a particular Web page, the auctioneer agent, acting on the user's behalf, sells sidebar space,[1] shown in Fig. 2, where relevant recommendations

[1] The currency used in our system is a notional one and is purely internal to the recommendation system. That is, the user does not receive any payments; the currency is simply a means of controlling the relative influence and impact of the constituent recommenders.

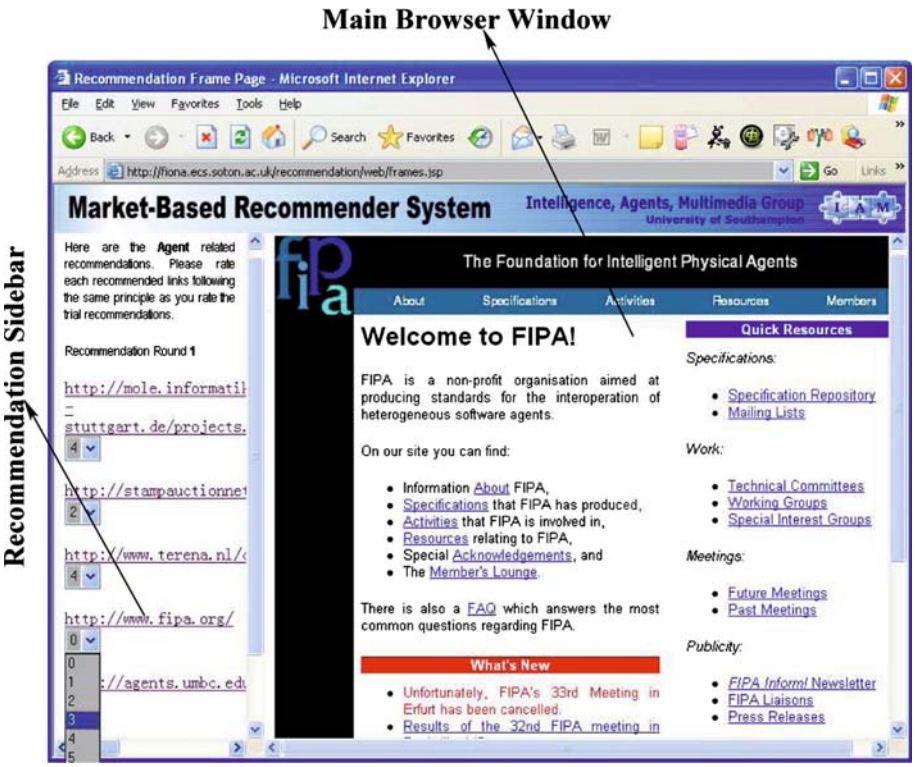

Fig. 2 Browser with recommendation sidebar

can be displayed. (In our case the sidebar has M slots and these are ordered in terms of decreasing relevance of the recommendation.) The recommender agents are keen to get their recommendations advertised in this space because they may receive a reward for so doing and they are assumed to be economically rational actors that seek to maximize their utility. Thus, each recommender agent identifies any items it believes are relevant to the current context, based on its own rating method, and associates a price with these items that reflects the amount it is prepared to pay to have that item presented to the user. This amount reflects the agent's confidence in the quality and appropriateness of its recommendation; the higher it believes the quality is, the more it will be willing to pay and the more the corresponding reward will be if the recommendation is deemed relevant by the user. The auctioneer agent then collects all the bids, ranks them in order of decreasing bid price, and displays the top M priced ones to the user. For those recommendations that are displayed, the corresponding agent pays the auctioneer the amount they bid (non-displayed bids incur no costs). If the user indicates that any of the recommendations are valuable to them, the agents that put them forward are rewarded in proportion to their bidding price and to the degree to which the user likes them. In this way, the recommending agents are incentivised to align the degree of importance they attach to their recommendations with what the user values; so, overall, the system is able to effectively coordinate the various recommender methods.

To demonstrate the suitability of this novel approach to recommender systems, our previous work carried out an analytical study and empirical evaluation of the system. Specifically, we established various economic properties of the marketplace such as its stability

and convergence; then we empirically verified the dynamic behavior of the system with simulated users and recommender methods that were given idealized interest profiles and that were assumed to make entirely consistent and rational choices. While these results were all encouraging, the real test is whether such a system actually works in practice with real users. Thus, in this paper, we report on the user evaluation of our market-based approach. In so doing, we advance the state of the art in the following ways. First, we show that our market-based approach is indeed capable of effectively coordinating multiple recommenders so that high quality recommendations are consistently placed in front of users. Second, we demonstrate that our system is capable of putting forward the best recommendations that are available from the constituent recommenders. Third, we show that a well coordinated ensemble of recommenders is capable of delivering superior recommendations than any of its constituent components. By so doing, we demonstrate that a market-based approach offers a powerful new paradigm for constructing recommender systems.

The remainder of the paper is structured as follows. Section 2 briefly outlines the design of the market mechanism and the bidding strategy of the recommender agents. Section 3 presents the metrics we used to evaluate our system and Section 4 details the user trial process. Section 5 then analyzes the results of the trials. Finally, Section 6 concludes.

2 The market mechanism design

In this section we detail the auction protocol we designed, the reward mechanism we established, and the bidding strategies of the individual agents.

2.1 The auction protocol

This section defines the auction protocol for managing the multiple recommending agents (as per Fig. 1). To ensure recommendations are provided in a timely and computationally efficient manner, we choose a *generalized first-price sealed-bid auction* in which all agents whose recommendations are shortlisted pay an amount equal to their valuation of the advertisement (meaning we have price differentiation[2]). We choose a sealed bid auction (in which agents will typically make a single bid) to minimize the time for running the auction and the amount of communication generated. We choose a first price auction with price differentiation because the relative ordering of the recommendations effects the likelihood of them being selected by the user. In particular, in the market, each information provider agent is keen to get their recommendations advertised to the user. Each agent has a valuation of the recommendation (which will be different for the different agents) and is willing to pay up to this amount to display its recommendations. When an agent gets its recommendations shortlisted, and therefore advertised to the user's browser, it has consumed the advertisement service provided by the recommender system. In return, it needs to pay an amount of credit (at the bidding price) to the system for each of its shortlisted items.

[2] If there is more than one item to be sold, the items can all be sold at the same price (called price uniformity) or they may be sold at different prices (called price differentiation). In this work, we exploit price differentiation because it differentiates recommendations so as to display them at different advertisement slots and it allows a seller to obtain the maximum possible profit. This approach has certain similarities to the sponsored keyword auctions that are now run by several search engines (although our work started independently and before these auctions were widely used). However, there are also a number of important differences; specifically, we use the auctions to personalise the recommendations to *specific individuals* and to provide feedback to the bidding agents (recommenders) so they can align their bidding with the preferences of the users.

In more detail, the market operates in the following manner. Each time the user browses a new page the auction is activated. In each such activation, the auctioneer agent calls for a number of bids (M which equals the number of recommendations being sought). Then each bidding agent submits up to M bids. After a fixed time, the auctioneer agent ranks all the bids it received by their bidding price, and directs the M bids with the highest prices to the user's browser sidebar (as shortlisted recommendations). Those bidding agents whose recommendations are shortlisted pay the auctioneer agent according to how much they bid. Those bidding agents whose recommendations are not shortlisted do not pay anything. The user may then follow up a number of the shortlisted recommendations in which case the agent that supplied them is rewarded.

More formally, the protocol for each auction round is defined in Fig. 3. It should be noted that: (i) function $GenerateBid\ (A_{bi}, rec_j, price_j)$ relates to the bidding strategy and will be discussed in Sect. 2.3; (ii) function *User Selects Recs* (SU) is the user making choices among the shortlisted recommendations; and (iii) function *Compute Reward* (b_h) concerns the reward mechanism and will be discussed in Sect. 2.2.

2.2 The reward mechanism

With the auction protocol in place, we now turn to the reward mechanism. According to our protocol, the user may select multiple recommendations from the shortlist. For each such user-selected recommendation, the suggesting agent is given a reward. In defining the *Compute Reward* function, our aim is to ensure that it is both Pareto efficient and social welfare maximizing (as motivated in [13]). Since the global objective is to shortlist the most valuable recommendations in decreasing order of relevance, *as perceived by the user*, we decided to reward the user-selected recommendations based on this feedback. Given this, the user-perceived quality (UPQ) for a given user for the set of N selected recommendations can be defined as Q_h ($h \in [1 \ldots N]$ and Q_h is a positive natural number that represents a user's ratings or preferences of the interesting recommendations). In practice, however, all user-selected recommendations are ordered in decreasing rank of UPQ such that $Q_1 \geqslant Q_2 \geqslant \cdots \geqslant Q_N$. Thus, Q_h denotes *the* hth *rewarded recommendation* (user-selected recommendation with the hth highest UPQ). To ensure different quality recommendations' bidding prices converge to different levels (so that our marketplace is able to differentiate recommendation qualities), we involve two other variables: P_h ($h \in [1 \ldots N]$) and P^*_m ($m \in [1 \ldots M]$). The former is the bidding price of the hth rewarded recommendation. The latter is the historical average bidding price of the mth shortlisted recommendation during the system's lifetime (note the bidding agents do not actually know this value). By this definition, P^*_m indicates the price for the mth advertisement displayed in the user's browser sidebar which is decided by the "invisible hand" (namely the market). With this information, we can define the reward to the hth rewarded recommendation as:

$$R_h = \delta \cdot Q_h \cdot P_{M+1} - \alpha \cdot |P^*_h - P_h| \tag{1}$$

where δ and α are two system coefficients ($\delta > 0$ and $\alpha > 1$) and P_{M+1} is the highest not shortlisted bid price (the detailed justification for this particular choice is given in [13]). The values of δ and α will depend upon the specifics of the application, but they need to be set at suitable values to ensure $R_h > P_h$ so that the rewarded agents can make profits. We base the reward on P_{M+1} (whose value is not known by the bidding agents) so that the market cannot easily be manipulated by the participants. This approach also reduces the possibility of bidding collusions because the reward is based on something that the rewarded agents are unaware of and cannot control.

The Variables:

S: the number of recommending agents ($S \gg 1$); — we assume the number of recommenders makes the number of recommendations sufficiently large with respect to the number of sidebar slots such that there is sufficient competition to make the marketplace operate efficiently.

$A_{b1}, A_{b2}, ..., A_{bS}$: S bidding agents;

A_B: complete set of bidding agents, i.e., $A_{b1}, A_{b2}, ..., A_{bS}$;

A_a: auctioneer agent;

A_u: user agent;

T_b: duration of the auction;

M: number of recommendations that A_u requests from A_a;

$b_{ij} = \langle A_{bi}, rec_j, price_j \rangle$: bid provided by A_{bi}, containing the j^{th} recommendation with bidding price $price_j$ ($i \in [1..S], j \in [1..M]$);

B^{ALL}: a set of bids which represents all bids submitted to A_a;

B^M: a set of bids which represents the shortlisted bids that will be recommended to A_u;

B^R: a set of bids which represents those selected by the user (and will be rewarded by A_a);

SU: a set of recommendations displayed in the user's sidebar (i.e. B^M ignoring the prices);

SU^R: a set of recommendations that are selected by the user (i.e. B^R ignoring the prices);

N: number of user-selected recommendations;

b_l, b_h: two bids for temporary use ($l, h \in [1..M]$);

R_h: reward to h^{th} user-selected recommendation.

The Algorithm:

$B^{ALL} = \phi$; $B^M = \phi$; $B^R = \phi$; *// system initialization*
$CallForBids(A_B, M, T_b)$; *// system calls for bids*
repeat during the duration of auction T_b
{
  $b_{ij} = GenerateBid(A_{bi}, url_j, price_j)$;
  $B^{ALL} = B^{ALL} \cup \{b_{ij}\}$;
}
for $l = 1$ **to** M **do** *// shortlist M highest bids*
{
  $b_l = FindBidWithLthTopPrice(B^{ALL}, l)$;
  $B^M = B^M \cup \{b_l\}$;
}
$SU = \{ \langle A_{bi}, url_j \rangle \mid \langle A_{bi}, url_j, price_j \rangle \in B^M \}$; *// the set of shortlisted URLs*
$SU^R = UserSelectsURLs(SU)$; *// user makes selection ($SU^R \subseteq SU$)*
$B^R = \{ \langle A_{bi}, url_j, price_j \rangle \mid \langle A_{bi}, url_j \rangle \in SU^R$ *and* $\langle A_{bi}, url_j, price_j \rangle \in B^M \}$;
$N = |B^R|$; *// the number of user selected items*
for $h = 1$ **to** N **do**
{ *// reward the user selected items*
  $b_h = FindHthBid(B^R, h)$;
  $R_h = ComputeReward(b_h)$;
}

Fig. 3 The auction protocol

2.3 Designing the agents' bidding strategies

In our marketplace, three kinds of information are revealed to a bidder with regards to a specific recommendation: (i) the score/relevance computed by its underlying algorithm that is making the recommendation (this is here termed its internal quality or INQ), (ii) this bidder's last bid price (P^{last}) and (iii) the previous rewards to this recommendation (a bidder actually knows the second piece of information). With this information, a rational bidder seeks to maximize its revenue by bidding sensibly for recommendations based on its knowledge of previous outcomes. Such bids can result in one of the following outcomes occurring: the bid is not shortlisted, it is shortlisted but not rewarded, or it is rewarded. With respect to a given INQ level, a bidder's strategy depends on the last outcome in the following way (again see [13] for a justification for these choices):

- *Bid Not Shortlisted*: The only way to increase revenue is to get the recommendation shortlisted. Therefore, the agent will increase its bidding price:

$$P^{next} = Y \cdot P^{last} \qquad (Y > 1)$$

- *Bid Shortlisted But Not Rewarded*: This means the agent overrated its INQ with respect to the UPQ and so the agent should decrease its price in subsequent rounds so as to lose less:

$$P^{next} = Z \cdot P^{last} \qquad (0 < Z < 1)$$

- *Bid Rewarded*: These agents have a good correlation between their INQ for a recommendation and that of the UPQ. Therefore, these agents have a chance of increasing their revenue. The profit made by the hth rewarded recommendation is:

$$\xi_h = \delta \cdot Q_h \cdot P_{M+1} - \ \alpha \cdot |P_h^* - P_h| - P_h$$

However, the agent is unaware of P_h^* (as per Sect. 2.2), so in practice it does not know whether ξ_h has been maximized. Hence, it must minimize $(\alpha \cdot |P_h^* - P_h| + P_h)$ so as to maximize ξ_h. Furthermore, the agent does not know whether P_h is higher or lower than P_h^*. In either case, however, the agent will definitely make a loss if P_h is not close to P_h^*. Therefore, we find that the hth rewarded agent can always be aware of whether its price is closer to or farther from the hth historical average market price, P_h^*, by adjusting its bidding prices (see [13] for the formal proof).

We have previously proved that a rational rewarded bidder will adjust its price to the corresponding average market price to maximize its profit [11]. Therefore, a rewarded agent's practical strategy with respect to certain rewarded recommendations is to bid in the following manner: whatever its current price is with respect to the historical average, when adjusting the bid price, if the adjustment results in making less profit, it indicates the action is wrong and $(P_h \pm \Delta P)$ is farther from P_h^*; if it results in making more profit, it indicates the action is right and $(P_h \pm \Delta P)$ is closer to P_h^*. This phenomenon is listed in Table 1 ($\Delta\xi$ represents the possible profit of the next bid compared to that of the current bid). In fact, Table 1 specifies the strategy for the rewarded agents: chasing the corresponding historical average market price. The actual value of ΔP will be defined in an application specific manner.

Table 1 Price adjustment and results

Current price	Adjustment	$\lvert P_h^* - P_h \rvert$	$\Delta\xi$
$P_h < P_h^*$	$+\Delta P$	$\searrow$	>0
	$-\Delta P$	$\nearrow$	<0
$P_h > P_h^*$	$+\Delta P$	$\nearrow$	<0
	$-\Delta P$	$\searrow$	>0

3 Evaluation metrics

In seeking to evaluate our system, the first step is to identify the properties that we want it to exhibit. In particular, we are interested in the following metrics (see [10] for a detailed justification for this choice):

1. *High Quality Recommendations*: The key feature of a recommender system is that it makes suggestions that the user finds valuable. To capture this, we define *high quality recommendations* as those that are rated highly by a user (see Fig. 1). Then we define two associated metrics: (i) a *qualified recommending round* and (ii) a *satisfied recommending round*. Specifically, with respect to a particular user visiting a particular Web page, a qualified round is an auction that results in at least one high quality recommendation being displayed in any advertisement slot of the recommendation sidebar, whereas a satisfied round is one in which at least one high quality recommendation is displayed in either of the first two advertisement slots. Thus, a satisfied round must be a qualified round, but a qualified round need not be a satisfied one.
2. *Effective Peak Performance*: If the marketplace is operating effectively, it should identify and promote the best recommendations. To check this, we compare the users' perceptions of the top-rated recommendation from our market-based system with that of the top rated items for each of the constituent methods. To do this, we define a metric called *peak performance*. Specifically, a constituent recommender's peak performance in a given auction is the user's rating of its highest price bid and the market-based recommender's is the rating of the item in the first position of the browser sidebar. Note that in the case of a constituent recommender that has no item shortlisted in a given auction, its peak performance is zero. Therefore, if our system is operating effectively, the market-based recommender's peak performance should be as high as that of the best of the constituent recommenders' for most auction rounds for most users (this we term *effective peak performance*). From this, we can evaluate how effective the marketplace is in picking out the best recommendations.
3. *No Dominant Method*: The key underpinning intuition of our market-based approach is that no individual recommendation method is likely to maximally satisfy all users in all situations. To determine whether this is indeed the case, we term the recommendations suggested by a constituent recommender and displayed in the browser sidebar its *output contributions*. Now, for a given user, it may be the case that one recommender makes the significant majority of output contributions and the others make very few. In such cases, we say that the recommender that contributes the majority of outputs *dominates* the marketplace. Such domination, with respect to a specific user, is not necessarily a bad thing (because it means the dominating recommender has learnt this user's interests more efficiently and therefore contributes more good recommendations than the others). However, it would be a problem if the same method dominates the entire user population

because it means that the marketplace essentially degenerates to that single dominant method and the rationale for having multiple constituent methods is no longer valid. Thus, if multiple coordinated methods are the best way forward, we would expect the different constituent recommenders to make broadly similar output contributions, given a broadly similar quality of recommendations at their disposal, when considered over the population of users.

With these metrics in place, we now outline the user trial process.

4 The user trials

In this section we detail the process we followed to perform the user trials. Our evaluation involved thirty-one participants who were academic staff, research fellows, and PhD students from the School of Electronics and Computer Science at the University of Southampton. Specifically, we drew mainly from members of the Intelligence Agents Multimedia (IAM) research group (http://www.iam.ecs.soton.ac.uk). These individuals have research interests in the areas of software agents, artificial intelligence, machine learning, knowledge technologies, game theory and Web technologies.[3]

First, we give some brief details of the actual set up of the market-based recommender system we have developed. Then we outline the set up phase of the trials where basic information is built up about the user population for use by the constituent recommender agents. Finally, we describe the activities involved in the actual operational data gathering part of the trials.

4.1 The market-based recommender system implementation

The marketplace is structured as described in Sect. 2 and an auction is run every time a user visits a new Web page. When a new auction is activated, each of the recommender agents submits M ($= 5$) sealed bids and the auctioneer ranks these in order of decreasing price. Recommendations that are valuable to a user are rewarded as described in Sect. 2. Our previous analytical work has proved that such a mechanism rewards the agents that can best align their bids with the users' interests. In this case, agents that over rate their recommendations by giving them an inflated price quickly lose revenue because they will pay high prices to get their items advertised, but they will receive a comparatively smaller reward. In contrast, agents that under value their recommendations, by giving them a deflated price, are less likely to get their recommendations displayed and so will fail to accumulate any reward. Within this regime, each constituent recommender agent has a distinct set of potential recommendations that it can put forward (i.e., there are no overlaps between the items that each recommender can put forward). It segments these into a number of rating levels (here 6) based on the quality of the recommendations as computed by its underlying ranking method. Now, each rating level is initially assigned an identical probability of making a recommendation. In the beginning, each agent also randomly selects M items from these internal rating levels with the same probability. After bidding and (potentially) receiving rewards, the agent computes how much revenue, on average, each of these rating levels are expected to make. From this, it updates the probability of making recommendations at each rating level, using a standard

[3] We acknowledge that this is a skewed population of users in that they are all highly computer literate. Nevertheless, we do not believe there is anything in our experimental set-up and analysis which means the results obtained would be different for a more general user population.

reinforcement learning strategy [12], so that the higher its expected revenue the higher the probability of it being chosen to make a recommendation. In this way, the recommender agent learns and adapts its bidding according to the user's preferences.

There are three broad types of recommendation method that are incorporated into the system:

- a content-based method that uses the similarity between the current document and those the user has previously indicated as being of interest;
- a collaborative method that uses the correlations between the user's interests and those of other users';
- a demographic method that uses the similarity between the available documents and the user's profile as represented by their keyword topics of interest.

Therefore, our constituent recommendation methods are each based on different similarity measures: document-to-document, user-to-user, and document-to-user. Each of the methods is represented as a separate agent acting as a recommender in the marketplace. These agents use well established versions of each of these methods since our focus is on the marketplace and how it can coordinate the methods, rather than in optimizing each of the individual methods themselves.

We now provide more details of each type of recommender:

- *The Content-Based Method*: This suggests recommendations based on the contents of the user's top rated documents. Therefore, in the trial set up phase, this method needs to learn something about documents that the user thinks are valuable (see Sect. 4.2 for more details). Once this has happened, this method computes the similarity between the current page being browsed and those potential recommendations that it could make by extracting the keywords with the highest term frequency from each document [8].[4] After experimenting with various numbers, we found that fifteen keywords represents a good tradeoff between computational tractability (storing more keywords is more resource intensive) and recommendation accuracy (storing fewer keywords leads to less accurate recommendations).
- *The Collaborative Method*: This suggests recommendations based on the similarity between the current user and other members of the population. The model we use here is based on [5] and involves the system putting forward recommendations that were highly rated by similar users. Here the similarity between users is obtained using Pearson correlation [2], in which each user is represented as a vector of their ratings on different interest topics and the similarity between two users is computed by the cosine of the two vectors. To make this method work effectively for our trials, we need to overcome the cold start problem.[5] Here we use a "collaboration via contents" technique to predict the likely rating of the source recommendations [9]. Thus, for each potential Web page selected for recommendation, a rating value is assigned by computing the number of keywords shared between the document and the user's interests (see the demographic method below for more details). Thus, when there are an insufficient number of similar users, it is still possible to predict their ratings using this method.

[4] To extract the most frequently occurring keywords from a Web document, a lookup table is used to filter out unimportant words that do not make sense in our context and need to be ignored (such as "a", "the", "in", "that" and "and") [3]. A stop-list technique, also taken from Middleton's work, is used to match different words with the same meaning. For example, "negotiation", "negotiations", "negotiating" and "negotiated" are tokenized into "negotiat" and are all deemed the same word.

[5] This happens when the first few individuals start to use the system and occurs because it is unlikely that any other users have similar interests (because the sample size is simply too small). In such cases, this method has no basis for putting forward recommendations.

- *The Demographic Method*:[6] Here we do not analyze people's characteristics in terms of traditional demographic measures, but rather by their research interests (since what we recommend are Web documents that are relevant to a particular set of topics of interest). Thus, we group people by characteristics of their topics of interest and match people to documents with relevant topics. We do not consider this method to be a content-based one, nor a collaborative one (although it does analyze the textual contents of documents and group people with similar interests). Specifically, it is not a content-based method because these compute similarities between documents and it is not a collaborative method because these compute similarities between people. In contrast, this method computes the similarity between the characteristics of people and the attributes of recommendations. Indeed, the main difference between this method and a typical demographic one lies in the fact that we use the browsing interest characteristics of people instead of the typical demographic ones. However, they both essentially match items to the group's common interests. For example, a group of people who share the interest topic of "machine learning" should all be interested in documents related to "reinforcement learning".

We now turn to the way in which the trial was set up.

4.2 The trial set up phase

The user trials were split into two stages (see Fig. 4). The first, dealt with in this subsection, was concerned with obtaining the information that was necessary for the constituent recommenders to operate. The second, dealt with in the next subsection, was concerned with the operational phase of the trial in which the performance of our market-based system was measured.

To ensure our results were not affected by any biases that might have occurred while the constituent recommenders were learning the users' interests, we went through an explicit user profiling stage that provided the necessary information that the three types of recommender agent needed in order to operate. When the system is actually deployed, such a stage will typically not be needed, but to ensure reproducible results within a short time frame we included such a stage here. We also limited the range of topics about which recommendations were made for similar reasons. In more detail, there were four steps in the trial set up stage (as per the upper part of Fig. 4). In the first step, a user selected the topic they wanted to investigate during the trial (step 1). The available *browsing topics* were: *software agents, automated negotiation and machine learning* (these were chosen as a result of an email survey about research topics of most interest to the user population) and each agent had over a 100 documents on each topic in its pool.[7]

In order to recommend good documents, the system needed to learn the users' interests. Thus, each constituent recommender needed to build a user profile as the basis to compute its recommendations. Since it is a difficult and complex process to precisely and automatically profile a user's interests [3] and because it was not the main focus of this work, we decided to do this in a relatively straightforward manner. From steps 2–4, three kinds of user interest

[6] A typical demographic method makes recommendations based on the demographic characteristics of people (such as age, gender and occupation) and groups people with similar characteristics [4]. Then, it analyzes the attributes of recommendations (such as textual descriptions or contents of books, colour or material of clothes and price of products), and, finally, matches people with certain characteristics to recommendations with suitable attributes.

[7] These documents were randomly allocated to each recommender before any ratings had taken place. Thus, on average, each agent had a broadly similar quality of base documents across each topic from which to make its recommendations.

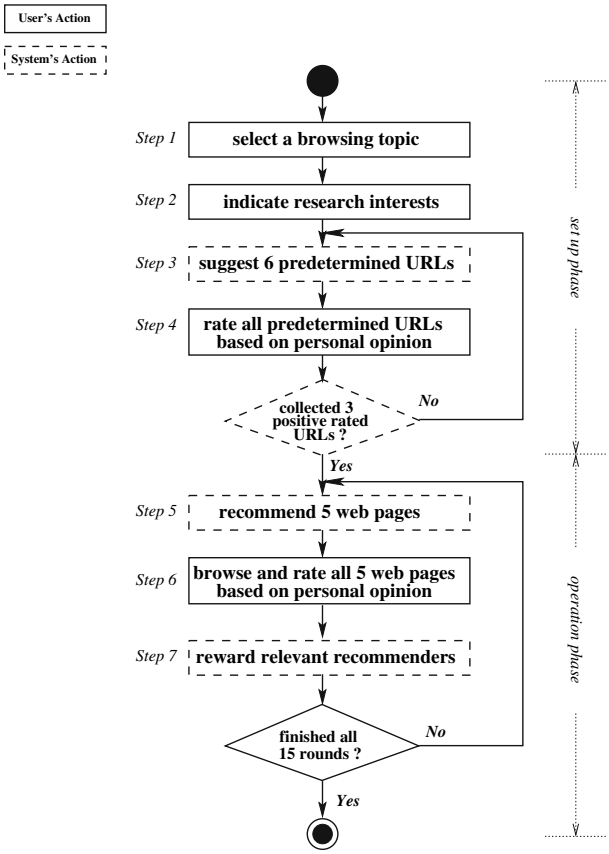

Fig. 4 The user trial process

profiles were built (one for each of the three methods because they computed their recommendations independently and used their user profiles in their own ways). In step 2, the user was required to rate a set of keywords that may be relevant to their research interests. These keywords were: *agents, biorobotics, artificial intelligence, machine learning, knowledge technologies, automated negotiation, auctions, markets, game theory, e-commerce, semantics, software engineering, information processing, distributed computing, grid computing, web services, networks, security, trust, mobility, ontologies and hypermedia*. A rating number was limited to the range between 0 and 5: where 0 indicated totally irrelevant, 1 indicated weakly relevant and 5 indicated perfectly relevant. Based on these ratings, the user profiles for the collaborative and demographic recommendation methods were built (see Sect. 4.1 for more details). To produce a profile for the content-based method, the system randomly selected six Web documents based on the user's chosen browsing topic and for each displayed their recommendation URLs in the browser sidebar (see step 3). We term these the *predetermined URLs*.[8] The user was then required to browse all these predetermined URLs and give each a rating according to their personal opinion (step 4). From these ratings, the

[8] The predetermined URLs were randomly selected from a separate recommendation pool from the three constituent recommenders that each had their own pools. Thus the four recommendation pools shared no common items.

content-based recommender collected a number of the most interesting documents and analyzed their contents to produce its user profile (which was represented as the five top-rated documents, where each document was represented as a vector of the fifteen most frequently appearing keywords). To capture the user's actual interests, at least three highly rated URLs were needed. If less than three were collected, this process was repeated until three were available. When more than five documents with the same highest ratings were collected, the latest one was added into the user's profile and the earliest one was removed (to place greater emphasis on the user's most recent opinions).

4.3 The trial operation phase

After the set up stage, a user entered the operational phase (the lower part of Fig. 4). In this stage, the market-based system presented five recommendations to the user (step 5) (and this was repeated fifteen times—which means a complete trial took between one and two hours). In each trial, the user examined all the recommendations presented to them and gave each a rating according to how relevant it was to their research interests (step 6).

For example, user 16 had a list of five interest topics `(agents, 3), (machine learning, 2), (auctions, 3), (markets, 4) and (information processing, 5)` (the numbers represent their relevance and the higher the number the more relevant the topic) and the other seventeen topics had zero relevance. This user had previously chosen "agents" as their browsing topic and was recommended two Web documents in this broad area. Specifically, one document was on a topic of "using market-based mechanisms to coordinate information agents" and the other was on "mobile agent security over the Internet". In this case, the user rated the former higher than the latter. This was because, besides agents, the former was related to markets and information processing which were also part of the user's interests, whereas the latter related to mobility and security which were not. For another example with respect to the same user, a third Web document was suggested on a topic of "agents and machine learning". In this case, the user preferred the first recommendation to this one because machine learning was less relevant than markets and information processing.

In short, a rating for a recommendation Web document is a user's personal opinion about how well the document relates to their research interests. Again the rating was limited to the range 0–5. We used five positive levels to specify recommendation quality because this number has previously been shown to be sufficient in differentiating users' preferences [1,5,6]. We assumed a user's rating of each recommendation was an absolute value that persisted throughout their trial. Thus, if a recommendation was rated by the user in an earlier time, they were not able to change its value if it was presented again. Having rated each of the five recommendations, the system rewarded the relevant constituent recommenders to assist their learning about the user's interests (step 7).

5 Evaluation

Having described the marketplace and outlined the trial process, we now report on the outcome of the trials with respect to the metrics defined in Sect. 3.

5.1 High quality recommendations

In the course of the trials, 436 effective recommendation rounds containing 2,180 recommendations were made to the 31 participants. Of these 436 rounds, 331 (75.9%) were

Table 2 Number of recommendations at different level and their distribution

Rating levels	"0"	"1"	"2"	"3"	"4"	"5"
Number of recommendations being made	329	268	388	493	419	283
Distribution (%)	15.1	12.3	17.8	22.6	19.2	13.0

qualified and 240 (55.0%) were satisfied. More specifically, the number of ratings at each of the levels is given in Table 2.[9] To contrast the qualities of the recommendations made by our system, they can be broadly classified into four categories: bad (completely irrelevant items, with rating 0), acceptable (items that have a degree of relevance, with rating 1), good (relevant items, with ratings 2 and 3), and very good (highly relevant items, with ratings 4 and 5). Thus, of all the recommendations made: 15.1% were bad, 12.3% were acceptable, 40.4% were good and 32.2% were very good. These raw numbers could naturally be improved, simply by improving the constituent recommendation algorithms or the user profiling process, but this is not the focus of our work.

In the context of this work, what was even more relevant was that our market-based system was putting forward the highest quality recommendations that were available to it. To determine whether this was the case, we needed to examine both the recommendations that were put forward and those that were not. This latter point is important because the system would not be operating effectively if very highly rated recommendations existed, but they were not put forward. To ascertain this, however, a given user had to go through the entire space of potential recommendations (of which there was over a hundred on each browsing topic for each agent) and assign each of them a rating. Thus we only did this for a sample of our trialists.

In more detail, Fig. 5 shows a typical example of these experiments from a randomly chosen user. Here, the horizontal axis represents the different rating levels and the vertical axis represents the number of recommendations. The white bars represent the numbers of available potential recommendations at each of the different rating levels. The light gray bars represent the numbers of items actually suggested by our system from the first to the fifth recommending round of the user's task, the dark gray bars the numbers suggested from the sixth to the tenth rounds, and the black bars those suggested from the eleventh to the fifteenth rounds. As can be seen, the white bars show that there were 18 recommendations (fifteen items with rating "4" and three with "5") that this user considered to be of high quality. Moreover, we can see that the numbers of these high quality recommendations has an overall tendency to increase over the recommending rounds. This indicates that our marketplace is able to effectively incentivise the constituent recommenders to learn the user's interests and to identify the best recommendations more frequently over time. From the numbers of recommendations made at rating levels "0" and "1", we can also see that our marketplace is able to deter such bad and weakly positive recommendations because the numbers of such recommendations have an overall tendency to decrease over time.

During this trial, from the first to the fifth rounds we found that there were four qualified recommending rounds and one of them was a satisfied recommending round; from the sixth to the tenth round, there were three qualified rounds and two of them were satisfied; and from the eleventh to the fifteenth round, there were five qualified rounds and four of them were satisfied (see Fig. 6). This meant that 80% of the first five rounds, 66.7% of the second five

[9] We believe this is a good result because the constituent recommenders are comparatively simple variations of the standard approaches and the user profiling process is also straightforward.

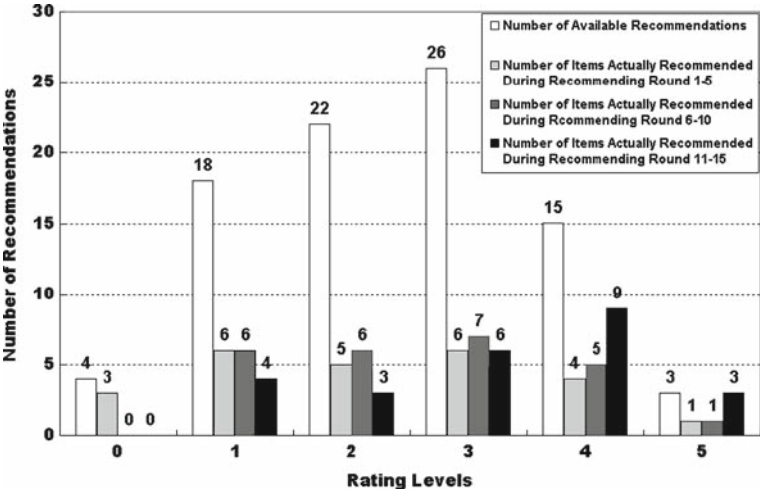

Fig. 5 Available recommendations versus actual recommended items

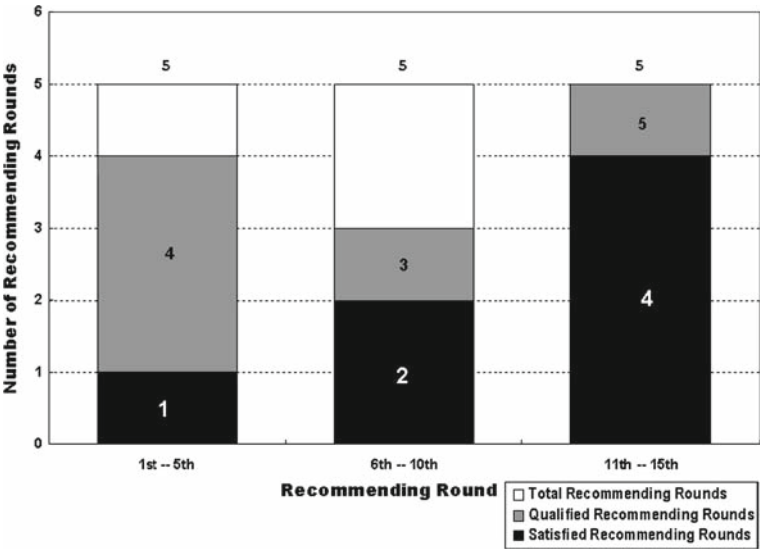

Fig. 6 Best recommendations identification for a given user

rounds and 100% of the last five rounds were qualified, whereas 20% of the first five rounds, 40% of the second five rounds and 80% of the last five rounds were satisfied. Therefore, both the qualified and the satisfied recommending rounds showed an overall tendency to increase.

When taken together, these results show that our marketplace is indeed able to identify the best recommendations and display them in the top positions of the recommendation sidebar quickly and frequently.

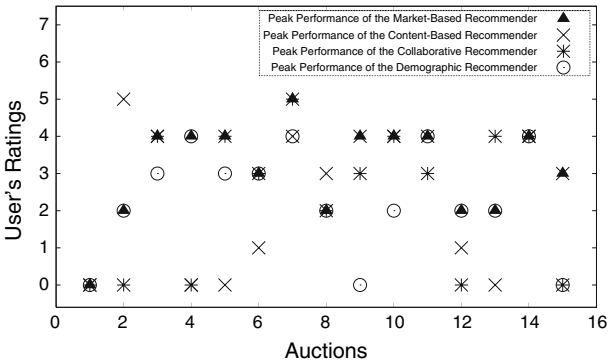

Fig. 7 Different recommenders' peak performances

5.2 Effective peak performance

To determine whether our market-based recommender's peak performance was indeed above that of all the constituent recommenders', we recorded their peak performance points for all users over all auction rounds. Specifically, Fig. 7 shows the marketplace's effective peak performance points versus those of the three constituent recommenders with respect to a particular participant. From this, we can see that the market-based recommender's effective peak performance points are at the first, third, fourth, fifth, sixth, seventh, ninth, tenth, eleventh, twelfth, fourteenth and fifteenth auction rounds. In the other three rounds, the recommendation displayed in the first slot of the sidebar was not the best of the constituent methods, but was the second best. This failure occurred because the constituent agents were still exploring their bidding prices to try and obtain the best fit with the user's interests. Overall, however, it is apparent that the marketplace's peak performance is, in most cases, above or equal to the best of the three constituent recommenders'.

To generalize this across the entire user population, we added up all the effective peak performance points for all the participants. From this, we observed that 66.4% of all the recommendation rounds for all users have their market-based recommender's peak performance as high as the best of the three constituent recommenders'. For the others rounds, which were primarily near the beginning of each trial, the market-based system picked the second best recommendation.

5.3 No dominant method

To evaluate the different constituent recommenders' actual contributions to the users, we recorded each method's output contribution for each user trial. We then computed the percentage of each constituent recommender's output contribution to each user over the complete trial. This information was recorded in Table 3 along with the standard deviation of the three methods' contributions with respect to each individual user. We were interested in the standard deviation in this context because it literally indicates the differences among the three methods' contributions (the bigger it was, the more likely a method was to dominate the marketplace). In this case, we chose the second deviation (15.28 with respect to user "2") as the criterion to differentiate whether or not domination occurs. This was because, with respect to a specific user, if the deviation was greater than or equal to this value, there must be one constituent recommender that contributes at least 2.5 times (see the second item in

Table 3 Different constituent recommenders' output contributions

User ID	Content-based recommender's output contribution	Collaborative recommender's output contribution	Demographic recommender's output contribution	Standard deviation of the three contributions
1	72 ★	20	8	34.02
2	50 ★	20	30	15.28
3	37.14	14.29	48.57 ★	17.46
4	28	53.33 ★	18.67	17.93
5	32	26.67	41.33	7.42
6	25.34	29.33	45.33	10.58
7	36	26.67	37.33	5.81
8	20	33.33	46.67	13.34
9	32	48	20	14.05
10	41.33	30.67	28	7.05
11	40.69	28.28	31.03	6.52
12	32	25.33	42.67	8.75
13	23.81	35.24	40.95	8.73
14	33.33	29.34	37.33	4.0
15	45.33	30.67	24	10.91
16	28	28	44	9.24
17	44	28	28	9.24
18	33.33	29.34	37.33	4.0
19	40	29.33	30.67	5.81
20	20	62.67 ★	17.33	25.44
21	32	33.33	34.67	1.34
22	22.67	30.67	46.66	12.22
23	22.67	40	37.33	9.33
24	40	22.67	37.33	9.33
25	54.67 ★	17.33	28	19.23
26	41.33	32	26.67	7.42
27	22.67	38.67	38.67	9.24
28	29.33	38.67	32	4.81
29	42.67	21.33	36	10.91
30	37.33	8	54.67 ★	23.59
31	29.33	44	26.67	9.33

A contribution with a ★ indicates its domination in the corresponding user trial

Table 3) more output contributions than another. This, we feel, is a reasonable, quantified view of dominance.

In more detail, in Table 3, the first column shows the anonymized identity of the participants. The second, third and fourth columns show, in percentage terms, the different constituent recommender's output contributions to each user. The last column shows the standard deviation of the three recommenders' contributions. From this, we can see that there were twenty-four user trials where no one method dominated, three trials dominated by the content-based recommender, and two trials dominated by the collaborative and the

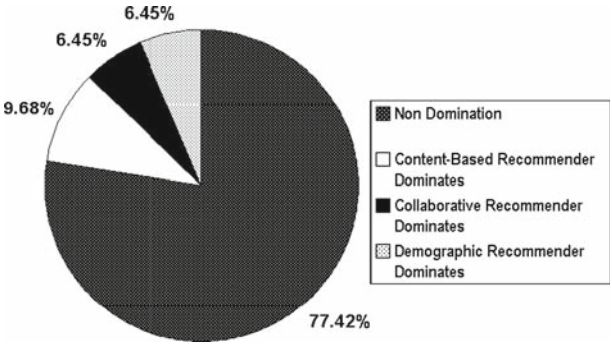

Fig. 8 Domination in the marketplace

demographic recommenders respectively (visually depicted in Fig. 8). This means that in most cases (77.42%) all three constituent recommenders made significant output contributions. From this, we conclude that the auction and reward mechanisms we have designed do not encourage domination in the marketplace.

The above analysis is based on individual users. However, we can also evaluate the overall contributions of the different recommenders to all users. This is important because it gives us an insight into the difference among the overall contributions of the different recommenders. To achieve this, we added up each individual recommender's output contributions to all users. This shows they contributed 35.1% (content-based), 30.8% (collaborative) and 34.1% (demographic) of the recommendations displayed to the users respectively. Again this indicates that, broadly speaking, each of the three constituent recommenders contributed about the same number of output contributions to the users, based on an equal quality of available recommendations, and so the marketplace is not biased towards any specific method.

6 Conclusions

This work has demonstrated the effectiveness and practicality of using a marketplace to coordinate multiple different recommendation agents. Based on the results of our user trials, we have demonstrated:

1. The marketplace works as an effective means of coordinating a variety of recommendation agents into a coherent overall system in which the best recommendations that are available, from whatever source, are placed in front of the user.
2. The market-based recommender is able to outperform any of the constituent recommenders in terms of placing high quality recommendations in the most prominent positions of the browser sidebar.

In sum, therefore, we have designed and implemented a market-based system that is able to combine multiple constituent recommendation methods into a coherent framework that is able to make high quality suggestions to users. We have determined its properties analytically and have demonstrated its performance through user trials. Given this, the next step is full scale deployment of the system.

Acknowledgements This research is funded in part by QinetiQ and the EPSRC Magnitude project (GR/N35816). We would like to extend thanks to Steve Hitchcock, Stuart Middleton and David De

Roure for their suggestions on designing the user evaluation of our recommender system. Additionally, thanks should also be given to Jingtao Yang for her help in developing the recommender system's web services.

References

1. Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems, 22*(1), 5–53.
2. Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2005). *Introduction to probability and statistics* (12th ed.). Thomson Brooks/Cole.
3. Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems, 22*(1), 54–88.
4. Montaner, M., Lopez, B., & Dela, J. L. (2003). A taxonomy of recommender agents on the internet. *Artificial Intelligence Review, 19*, 285–330.
5. Pazzani, M. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review, 13*(5-6), 393–408.
6. Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P., & Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. In *Proc. of ACM 1994 Conf. on Computer Supported Cooperative Work* (pp. 175–186). Chapel Hill, North Carolina.
7. Resnick, P., & Varian, H. R. (1997). Recommender Systems. *Commun. of the ACM, 40*(3), 56–58.
8. Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Addison-Wesley, Reading, Massachusetts. ISBN 0-201-12227-8.
9. Sarwar, B. M., Konstan, J. A., Borchers, A., Herlocker, J., Miller, B., & Riedl, J. (1998). Using filtering agents to improve prediction quality in the grouplens research collaborative filtering system. In *Proc. of the 1998 ACM Conf. on Computer Supported Cooperative Work* (pp. 345–354). Seattle, US.
10. Wei, Y. Z. (2005). *A Market-based approach to recommender systems*. PhD thesis, School of Electronics and Computer Science, University of Southampton.
11. Wei, Y. Z., Moreau, L., & Jennings, N. R. (2003). Recommender systems: A market-based design. In *Proc. of the 2nd International Joint Conf. on Autonomous Agents and Multi Agent Systems (AAMAS03)* (pp. 600–607). Melbourne, Australia: ACM Press.
12. Wei, Y. Z., Moreau, L., & Jennings, N. R. (2005). Learning users' interests by quality classification in market-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering, 17*(12), 1678–1688.
13. Wei, Y. Z., Moreau, L., & Jennings, N. R. (2005). A market-based approach to recommender systems. *ACM Transactions on Information Systems, 23*(3), 227–266.