J Ambient Intell Human Comput (2014) 5:105–110
DOI 10.1007/s12652-012-0138-7

ORIGINAL RESEARCH

Adaptive information retrieval system via modelling user behaviour

**Saeedeh Maleki-Dizaji · Jawed Siddiqi ·
Yasaman Soltan-Zadeh · Fazilatur Rahman**

Received: 4 December 2011 / Accepted: 14 May 2012 / Published online: 21 June 2012

Abstract There has been an exponential growth in the volume and variety of information available on the Internet, similarly there has been a significant demand from users' for accurate information that matches their interests, however, the two are often incompatible because of the effectiveness of retrieving the exact information the user requires. This paper addresses this problem with an adaptive agent-based modelling approach that relies on evolutionary user-modelling. The proposed information retrieval system learns user needs from user-provided relevance feedback. It is proposed that retrieval effectiveness can be improved by applying computational intelligence techniques for modelling information needs, through interactive reinforcement learning. The method combines qualitative (subjective) user relevance feedback with quantitative (algorithmic) measures of the relevance of retrieved documents. An adaptive information retrieval system is developed whose retrieval effectiveness is evaluated using traditional precision and recall.

Keywords User information needs modelling · Interactive evolutionary learning · Adaptive information retrieval

S. Maleki-Dizaji
Department of Computer Science,
The University of Sheffield, Sheffield S1 4DP, UK
e-mail: s.maleki-dizaji@dcs.shef.ac.uk

J. Siddiqi (✉) · F. Rahman
C3RI, Sheffield Hallam University, Sheffield S1 1WB, UK
e-mail: J.I.Siddiqi@shu.ac.uk

Y. Soltan-Zadeh
School of Management, Royal Holloway University of London,
Egham Hill, Egham TW20 0EX, UK
e-mail: pqtm497@live.rhul.ac.uk

1 Introduction

The emergence of the World Wide Web (WWW) has resulted in access to a vast and exponentially growing, unstructured and dynamic network of information sources. However, it can now be argued that the volume of information available on-line is becoming a hindrance to effective information retrieval. A number of factors affect information retrieval effectiveness. The most significant three for our purposes are as follows. First, users often do not present search queries in the form that optimally represents their information needs. Second, the measure of a document's relevance is often highly subjective between different users. Third, information sources might contain heterogeneous documents, in multiple formats and the representation of documents is not unified. It is timely and necessary, therefore, to build new tools aimed at helping users retrieve documents that satisfy their information needs accurately and efficiently. An optimal information retrieval (IR) system is one which would be able to obtain from an information source only those documents that are relevant to a user's information needs, while at the same time excluding documents that are non-relevant.

One traditional way of adapting the information access process to user's information needs is relevance feedback. Relevance feedback aims to learn the user's requirements during a search session and by implicit or explicit user involvement adapt the behaviour of the retrieval process. In principle relevance feedback is a personalisation method; however, it only considers the short history of user behaviour in its modelling. To incorporate user's preferences and improve the information access performance, modelling the long term history of user's behaviour and providing a facility to augment the model manually can be more effective than a simple relevance feedback.

The above requirements, in combination with the other common information retrieval features, such as, document representation and ranking, lead to a system of relatively high complexity, because various time consuming tasks need to be carried out.

We propose a solution to this problem based on a multi-agent modelling approach. Agent-based modelling treats each individual component of a system as a single entity (or agent) obeying its own pre-defined rules and reacting to its environment agents accordingly thereby making the system more flexible and reliable. User modelling for information retrieval is done through genetic algorithm (GA) to evolve and adapt query vectors that are representative models of the user's information needs.

The rest of the paper is organised as follows: Sect. 1.1 presents background information and overview of the related work. Section 2 describes the architecture of the adaptive IRS and Sect. 3 presents the results of the experiments. Section 4 concludes the paper.

1.1 Background and related work

Genetic algorithms are a class of evolutionary algorithms that represent the solution candidates as a vector of bits or numbers. Similar to other evolutionary algorithms many individual solutions are randomly generated to form an initial population. Some of these individuals are selected through a fitness-based process to breed a new generation. The next step is to generate the next generation of candidate solutions from those selected through genetic operators, such as: crossover, and/or mutation. Crossover is an operator which forms two children by combining parts of two parents and mutation is an operator that forms a new individual by making minor changes to the parent's gene. This generational process is repeated until a termination condition has been reached.

Evolutionary computation and genetic algorithms have gained more attention in recent years and have grown to many applications in different areas of artificial intelligence, information retrieval (Fan et al. 2009; Lopez-Herrera et al. 2009; Torres et al. 2009), text mining (Phua et al. 2010; Alcala-Fdez et al. 2011) and natural language processing (Kao and Poteet 2005; Atkinson and Matamala 2009). Freitas (2008) has discussed the use of evolutionary algorithms, particularly genetic algorithms and genetic programming, in data mining and knowledge discovery.

An area of information retrieval, which is being approached by evolutionary algorithms, is query expansion and reformulating queries in order to improve the retrieval quality based on the user behaviour or other aspects of the retrieval process. Vrajitoru (1998) introduces a new crossover operation in genetic algorithm specifically for creating new queries.

In (Araujo et al. 2010), an evolutionary algorithm is employed to combine clauses to reformulate a user query in order to improve the results of a similar search. The study starts with a review of the query expansion algorithms and discusses the negative effects of term correlation used in query expansion. Then it combines the query clauses with genetic algorithm in order to create a method to improve the result of stemming by reformulating the user query. In contrast to our work this study has ignored the user and pseudo relevance feedback as well as any re-weighting formulas. Instead it has focused on the term dependencies, their occurrences in simple experiments and their effects on performance.

Araujo and Pérez-Iglesias (2010) proposed a way to train a classifier for query expansion of too short or unspecific queries. In this study, the user's relevance judgments on a document set are used as fitness function for the genetic algorithm to train the classifier to identify distinguished terms for query expansion. The authors conclude that the genetic algorithm training can improve the query expansion quality. Their main focus is on a pseudo relevance feedback method and genetic algorithm is used to build a set of suitable terms for query expansion from the top documents of the initial ranked list.

Loia et al. (2007) have used the collaboration between an agent-based parsing activity and a user-based suggestion method to reveal the relevance/similarity among the pages they have crawled. An improved fuzzy clustering algorithm is used to perform a locally personalized classification based on user's point of view. The user behaviour and preferences that are extracted during user's navigation are used to present the personalised clusters. These clusters enable agent-based spidering to mining new pages and present them to the user as prototype pages.

Ganzha et al. (2010) have studied whether combining the result of several sources results better. They have applied their theory to three main algorithms: game theory, auction-based approach and consensus method. However, each algorithm depends on a different part of the process, therefore instead of a synergy between different algorithms to recommend one specific set of answers they suggest different results which may be in conflict with each other.

Yu and Jeon (2010) have proposed a context-aware recommender system. The system uses the user's history and current context to filter the content-based information in order to provide preferable items to the user.

Li et al. (2009) have implemented an intelligent assistant to do spam filtering task based on what it learns from user behaviour.

The agent-based modelling in this work, on the other hand, is employed to improve the retrieval performance based on the feedback provided by the user.

2 Characterisation, use case and architecture of an adaptive IRS

Adaptive user modelling techniques produce *Adaptive* systems that can learn something about each individual user and adapt their behaviour to monitor the user's activity pattern. They automatically adjust the interface or content provided by the system to accommodate user differences and changes such as user: skills, knowledge and preferences (Chen et al. 1998). For our purposes an Adaptive Information Retrieval System (IRS) can be characterised as shown in Fig. 1 along with the following assumptions:

- A program does (implicit or explicit) user modelling provided it can change its behaviour based on something related to the user in information filtering component(IF) component.
- A user model contains all information that the system knows about the user. It is generally initialized either with default values or by querying the user.
- Thereafter, it is maintained by the system, although the user may be able to review and edit their profile through user model browser (UM Browser).

The user modelling for information retrieval proposed in this paper, applies an evolutionary, genetic algorithm (GA) to evolve and adapt query vectors that are representative models of the user's information needs which is a fixed number of keywords that can change through GA operations; therefore, the exact significance and relationship between features in determining the relevance of retrieved documents is not explicit, but evolves under the GA (Goldberg 1989). In our research, the vectors are expressed as keyword terms and associated weights. Search vector are used to retrieve documents and compete against other search vectors to asses relevance by matching with the user information needs. In our research in order to effect adaptation of the user model within a reasonably short time, we have assumed that user information needs are stochastic but non-transient. In other words, the information needs vary in a subjective non-deterministic manner between users, but they do not change rapidly over time. In this way, a GA can evolve a model for different information needs.

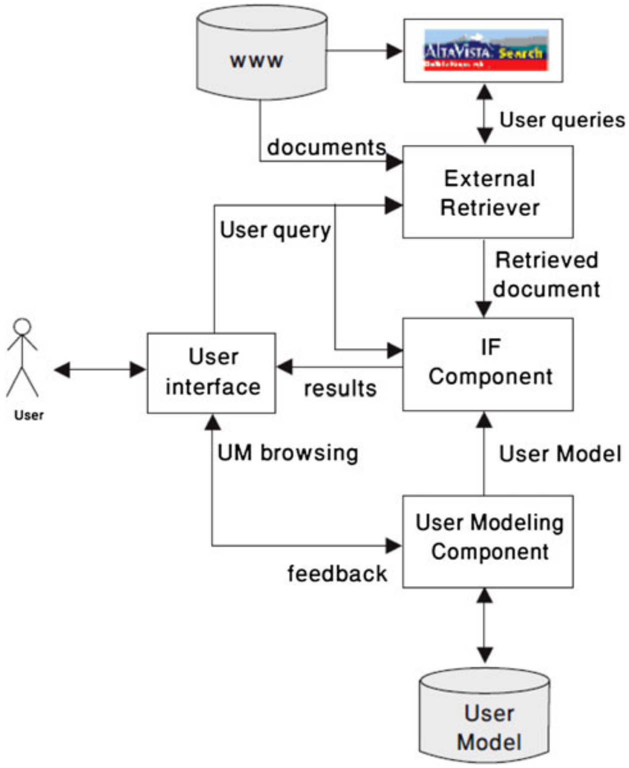


Fig. 1 Adaptive information retrieval system

The key role for GA in user-needs modelling is to continuously modify the representation of user needs. In our research we have based this on quantitative and qualitative relevance metrics. On one hand, a quantitative (algorithmic) metric is given by the similarity between the user-model chromosome[1] and the documents retrieved using the chromosome. On the other hand, each of the retrieved documents is given a qualitative assessment, interactively by the user. These two measures are then combined through a fuzzy inference system to derive an overall parameter that is used to adjust the "fitness of use" of competing information needs models. This is a novel learning approach we have termed evolutionary interactive reinforcement learning (EIRL).

The fuzzy inference system used to adjust the fitness of competing information needs model is a rule-based system that uses the similarity between a search vector and a retrieved document, and the user feedback to derive the required fitness modification for the search vector. These rules are, in general, heuristic but were fine-tuned by experimenting with the system. The underlying philosophy of the rules is to reward those vectors, which retrieve documents that the user judges to be relevant to his or her needs, and penalize those the user judges to be irrelevant. Thus, if the user judges a document to be relevant then the fitness of the search vector used for retrieval of the document should be increased, and especially more so if the algorithmic similarity measure is low. Conversely, if the user feedback is poor (not relevant) but, the algorithmic similarity between query and documents are high then the fitness of the chromosome should be reduced significantly.

The proposed adaptive IRS has been developed on a multi-agent paradigm to represent the different typical

[1] In evolutionary computation the data structure of the individual used for breeding is called *genome* and a *chromosome* is a vector-based genome.

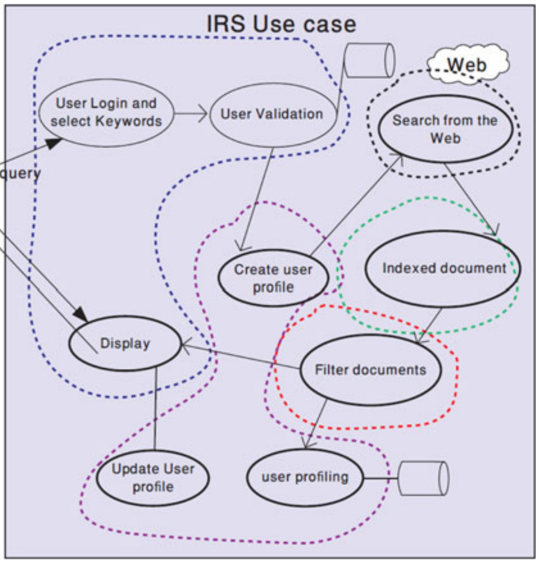


Fig. 2 The IRS use case

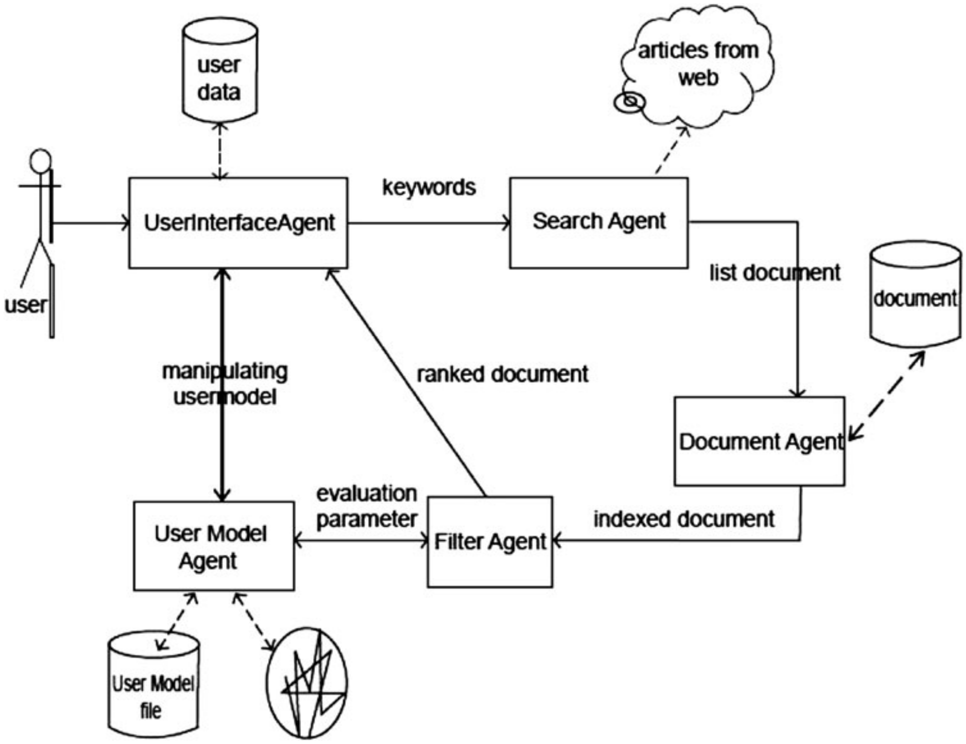
activities, including document representation, query formulation, user needs modelling and filtering and user-needs model reinforcement. The requirements of the system are represented using use case models, as shown in Fig. 2, by grouping those use cases into five independent task categories, which are assigned to agents to represent the IRS solution: user interface agent, search agent, document agent, filter agent and user model agent.

The agent system architecture is shown in Fig. 3. The *search agent* acts as a meta-search tool for any Internet search engines using the keywords to retrieve documents, which are then passed to the *document agent* which in turn indexes the documents using normalised keywords. As a result, the highest indexed documents are sent to the *filter agent* who ranks the indexed document, according to the user model in order to increase the precision. The *user interface agent* allows the user to evaluate the relevance of the ranked documents, by giving a score to each document in the form of fuzzy values modifying user-needs. In order to affect a perpetually evolving user model, the *user model agent* maintains a population of competing models, which evolve using genetic algorithm.

3 Evaluation of the adaptive IRS

Evaluation of any IR system calls upon examination of many issues including human computer interaction, usability and applicability of the system. In this research we focus on number of relevant documents retrieved with respect to a query which is known as retrieval effectiveness; this expresses how well the produced output satisfies a user's information need. The common performance indicators of retrieval effectiveness of IR system are recall and precision. Both indicators can be based on the user's subjective relevance assessments following the retrieval process. Recall measures the completeness of the output,

Fig. 3 Overall IR system architecture



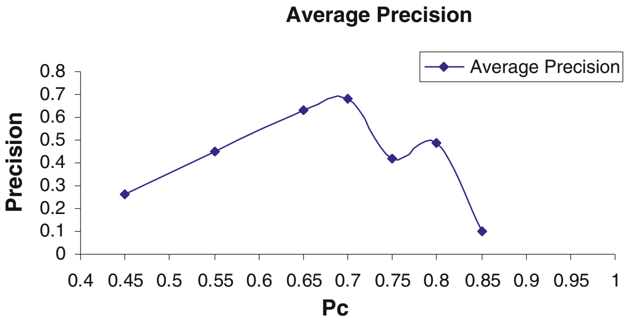


Fig. 4 Mean precision performance against different probability of crossover value

which is the ability of the system to retrieve all relevant information. Precision measures the relevance of the output, in other words, the ability of the system to reject irrelevant material. A good IR system should exhibit both high recall and high precision. There were two stages of the evaluation. The first stage was to evaluate the performance of the proposed evolutionary learning technique. This involves carrying out several experiments with the IR system, in order to fix the optimal GA parameters for the most efficient learning.

Figure 4 shows that low values of the crossover parameter do not correspond to significant improvement in learning because at low crossover values, fewer new search vectors are introduced to the population resulting in a longer time for the system to improve its performance. Conversely, too high a value for the crossover parameter results in introducing new vectors too quickly, which causes the system to change the population of user models more quickly and more randomly, regardless of user relevance feedback. The best performance of the system is given by medium values between 0.6 and 0.7 for crossover probability.

The second stage was to carry out interactive retrieval sessions with the different users. The experiments were carried using five PhD students in the area of Computer Science and Information Systems in the School of Computing and Management Sciences at Sheffield Hallam University, United Kingdom. The evaluated documents were obtained from the Bath Information and Data Services (BIDS) (http://www.bids.ac.uk). From the areas of the information needs of the assessors, 300 documents were selected and another 100 noisy documents, which had some common keywords but the contents were not relevant to the users' information needs, were added. The results shown in Figs. 5 and 6 are the mean values for the five different information needs. Each iteration represents a user search task, the total number of search repeat is 20.

The figures show the comparative information retrieval effectiveness of the proposed evolutionary learning IR system against a conventional relevance feedback (RF) technique. The latter is a technique that allows interactive reformulation of search query using information gained from retrieved documents known to be relevant (Salton and Buckley 1998). RF applies query reformulation where terms from relevant documents are added, and terms from non-relevant documents are removed from the query vector. Figures 5 and 6 show the average precision and recall for both techniques. It can be seen that overall the performance of the evolutionary approach EIRL is higher than conventional RF. However, it can also be observed that in the case of EIRL there are large fluctuations in performance over the period of user interaction with the IR system, which are attributable to the probabilistic nature of evolutionary learning. However, a statistical analysis was carried out to determine the significance of any differences between the two methods by comparing average precision values, across all queries using a paired *t* test (t = 2.10 and p = 0.032). Using a Confidence level of 0.05 from the statistical analysis we can claim that EIRL performs better than RF.

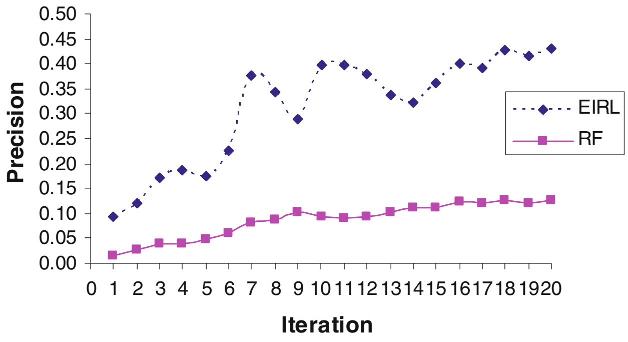


Fig. 5 Average precision for all queries

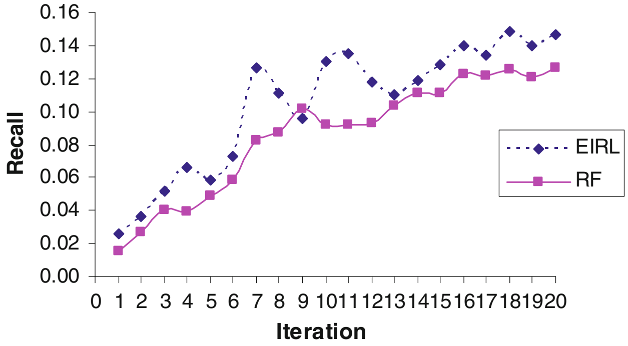


Fig. 6 Average recall for all queries

4 Conclusion

In conclusion we have described a novel approach for improving document retrieval effectiveness by combining fuzzy relevance feedback and evolutionary reinforcement learning. This approach has been evaluated especially for applications where user information needs are subjective but relatively static, and hence can be accurately modelled over a short period of time. Results obtained in this study suggest that the proposed approach, in general, performs better than conventional relevance feedback. Previous studies (Vrajitoru 1998) did not show any improvement over conventional relevance feedback when using an evolutionary approach for user modelling. This, however, can be attributed to the fact that they used binary coded genetic algorithms to represent the presence or absence of keywords. The proposed approach encodes chromosomes as keywords and their weights to imply their significance to user information needs profiles. The improvement in retrieval effectiveness, it is argued, is achieved by on-line reinforcement learning through interaction with users. Human interactive reinforcement provides a direct evaluation of the relevance of documents, namely, user preference that cannot be expressed by any analytical fitness function. This results in a user model that is, in fact, optimised by the user. The result also indicated that, in most cases, the maximum value for retrieval precision was reached in about ten generations, which suggests that learning can achieved in a relatively short period of interaction. This is desirable in order that specialization to user information needs is not a time-consuming exercise.

References

Alcala-Fdez J, Fernandez J, Luengo J, Derrac J, Garcia S, Sanchez L, Herrera F (2011) KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. Multiple Valued Logic Soft Comput 17(2–3):255–287

Araujo L, Pérez-Iglesias J (2010) Training a classifier for the selection of good query expansion terms with a genetic algorithm. In: IEEE Congress on Evolutionary Computation, pp 1–8

Araujo L, Zaragoza H, Pérez-Agüera JR, Pérez-Iglesias J (2010) Structure of morphologically expanded queries: a genetic algorithm approach. Data Knowl Eng 69:279–289

Atkinson J, Matamala J (2009) Chunking natural language texts using evolutionary methods. SGAI Conf., pp 277–290

Chen H, Shankaranarayanan G, She L, Iyer A (1998) A machine learning approach to inductive query by examples. J Am Soc Inf Sci 49(8):693–705

Fan W, Pathak P, Zhou M (2009) Genetic-based approaches in ranking function discovery and optimization in information retrieval—a framework. Decis Support Syst 47(4):398–407

Freitas AA (2008) A review of evolutionary algorithms for data mining. In: Soft computing for knowledge discovery and data mining. Springer, Berlin, pp 79–111

Ganzha M And, Paprzycki M, Stadnik J (2010) Combining information from multiple search engines—preliminary comparison. Inf Sci 180(10):1908–1923

Goldberg DE (1989) Genetic algorithms in search optimization and machine Learning. Addison-Wesley, Boston

Kao A, Poteet S (2005) Text mining and natural language processing: introduction for the special issue. SIGKDD Explor 7(1):1–2

Li W, Zhong N, Yao Y, Liu J (2009) An operable email based intelligent personal assistant. World Wide Web 12(2):125–147

Loia V, Pedrycz W, Senatore S, Sessa MI (2007) Interactive knowledge management for agent-assisted web navigation. Int J Intell Syst 22(10):1101–1122

Lopez-Herrera AG, Herrera-Viedma E, Herrera F (2009) Applying multi-objective evolutionary algorithms to the automatic learning of extended Boolean queries in fuzzy ordinal linguistic information retrieval systems. Fuzzy Sets Syst 160(15):2192–2205 (Elsevier)

Phua C, Lee VCS, Smith-Miles K, Gayler RW (2010) A comprehensive survey of data mining-based fraud detection research. CoRR. abs/1009.6119

Salton G, Buckley C (1998) Term weighting approaches in automatic text retrieval. Inf Process Manage 24:513–523

Torres RS, Falcao AX, Goncalves MA, Papa JP, Zhang B, Fan W, Fox EA (2009) A genetic programming framework for content-based image retrieval. Pattern Recognit 42(2):283–292 (Elsevier)

Vrajitoru D (1998) Crossover improvement for the genetic algorithm in information retrieval. Inf Process Manage 34(4):405–415

Yu J, Jeon M (2010) A context-aware intelligent recommender system in ubiquitous environment. In: 10th IASTED international conference on artificial intelligence and applications, pp 229–234