ASIAN JOURNAL OF INFORMATION TECHNOLOGY
MEDWELL PUBLICATIONS

Implementing Multimedia Information Retrieval using Memory-Based Collaborative Filtering

[1]Imeh Umoren, [2]Onukwugha Gilean and [1]Juliet Odii
[1]*Department of Computer Science, Akwa Ibom State University, Mkpat-Enin, Nigeria*
[2]*Department of Computer Science, Federal University of Technology Owerri, Owerri, Nigeria*

Key words: Intelligent agents, Collaborative Filtering (CF), memory-based CF and Jaccard similarity algorithm

Corresponding Author:
Imeh Umoren
Department of Computer Science, Akwa Ibom State University, Mkpat-Enin, Nigeria

Page No.: 60-76
Volume: 20, Issue 2, 2021
ISSN: 1682-3915
Asian Journal of Information Technology

Abstract: As the amount of information available to users on the internet increases geometrically, several approaches are required to assist the user in finding and retrieving relevant information. Intelligent agents with the capacity to learn user's profile towards efficient sentiment analysis are one solution to this problem. Collaborative Filtering (CF) is one of the most successful recommended approaches used in academia and industry for making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaboration). This work applies Memory-Based CF using Jaccard similarity algorithm in electronic commerce to develop a recommendation system for analyzing user data and extracting user information for accurate predictions of user preferences based on user's behavior in a Business-to-Consumer (B2C) E-commerce store. The results outcome indicates that CF as a classical method of information retrieval can be used in helping people deal with information overload as the technique reduces the time spent searching for relevant information and also increases the accuracy of retrieval. Furthermore, the results from predictions of user's interests through recommendation lists are useful for enhance customer's loyalty and higher marketing rates.

INTRODUCTION

The society is undergoing rapid transformation in almost all aspects due to technological advancement. Online transactions, information gathering by search engines and social networking on the internet are trending. Many transactions and interactions are stored electronically, thereby giving researchers the opportunity to study socio-economic and techno-social characteristics relying on the availability of massive datasets[1]. Information Retrieval (IR) deals with the representation, storage and retrieval of unstructured data. In the past only text was considered. Today, the evolution of multimedia databases and the web have given new interest to Multimedia Information Retrieval (MIR) systems. MIR is referred to the search for knowledge in all the digital media forms which can also include across multiple independent information attributes within a single data-stream. Thus, web-image search, news video retrieval and music retrieval are all different manifestations of MIR. A

MIR system can store and retrieve attributes, text, images, voice and music and video contents. Retrieval is based on the understanding of the content of documents and of their components. Retrieval can be broadly classified on the basis of the aspect of documents that each of them addresses.

Thus, retrieval can be based on syntactic similarity, on semantics, on structure and on profile. According to Faloutsos[2], form-based retrieval includes all similarity-based image retrieval methods. On the contrary, semantic-based retrieval methods rely on symbolic representations of the meaning of documents, i.e., descriptions formulated in some suitable knowledge representation language, spelling out the truth conditions of the involved document[3].

Furthermore, accuracy and speed as goals of retrieval must ensure that documents that the user expects in the answer to his query are retrieved and has to be fast. As the number and types of MIR providers increase steadily, the central concern of MIR will be that given a collection of multimedia documents (i.e., a complex information object, with components of different kinds, such as text, images, video and sound, allin digital form), find those that are relevant to information need of the user. Real-life applications for multimedia information retrieval systems abound in medical databases (X-rays, CT, MRI Scans), financial stock markets, criminal investigations (suspects, fingerprints), personal archives (text, color images) and scientific databases such as sensor and surveillance data, weather, geological and environmental data.

Information retrieval is not an exact process of searching as two documents are never identical but can be similar. Thus, searching must be approximate. The effectiveness of IR depends on the types and correctness of descriptions used, types of queries allowed, user uncertainty as to what he is looking for and the efficiency of search techniques. Text retrieval is a well-researched area and most systems work with text and attributes. Often, descriptions are extracted and stored either in the same or separate storage with the documents. The problem is that text descriptions for images and video contained in documents are not available, e.g., the text in a web site is not always descriptive of every particular image contained in the web site. Human annotations and feature extraction are the two approaches to text descriptions for images and video contained in documents. The former approach is inconsistent or subjective, time consuming, expensive, difficult for large databases and retrieval will fail if queries are formulated using different keywords or descriptions. Queries address the stored descriptions rather than the documents themselves. In the later approach, features are extracted from audio, image and video. It is a cheaper and replicable approach with consistent descriptions but gives inexact low level features (patterns, colors, etc). However,

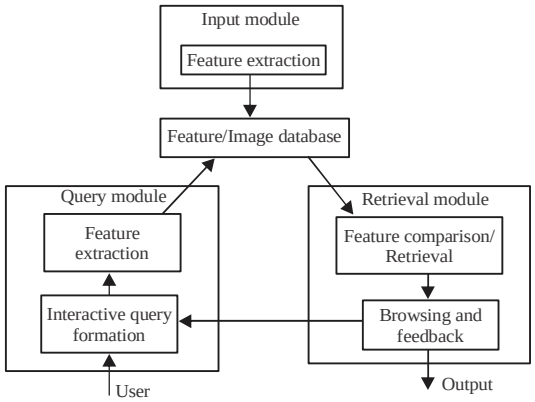

Fig. 1: The architecture of an information retrieval system

different techniques are used for different data. The architecture of an Information Retrieval System (IRS) for images is shown in Fig. 1.

In approximate IR, a query is specified and all documents up to a pre-specified degree of similarity are retrieved and presented to the user ordered by similarity. The two common types of similarity queries are the range queries and the nearest-neighbor queries. The range queries retrieve all documents up to a distance "threshold" while the nearest-neighbor queries retrieve the best matches. In deciding whether two documents are similar, distance similarity holds that the lower the distance, the more similar the documents and the higher the similarity, the more similar the documents are. For a successful retrieval, the more accurate the descriptions, the more accurate the retrieval will be. The focus of this work, therefore, is to use Jaccard similarity algorithm for collaborative filtering or automatic prediction of a user's interest by collecting preferences or tastes information from many users. The application is useful in the area of E-commerce for an item to item match in an E-store.

STATEMENT OF THE PROBLEM

The two fundamental necessities for a multimedia information retrieval system are searching for a particular media item and browsing and summarizing a media collection. In searching for a particular media item, the current systems have significant limitations such as an inability to understand a wide user vocabulary and the user's satisfaction level and there is non-existence of credible representative real-world test sets for evaluation or even benchmarking measures which are clearly correlated with user satisfaction. In general, current systems have not yet had significant impact on society due to an inability to bridge the semantic gap between computers and humans. In human-centered computing, the main idea is to satisfy the user and allow the user to

make queries in their own terminology. However, since, the primary goal is to provide effective browsing and search tools for the user, it is clear that the design of the systems should be human-centric.

Furthermore, the capacity of collaborative filtering system is yet to be fully harnessed. This is due to the fact that collaborative filtering systems still depend on only the analysis of information from the past activities of specific user, or the history of other users of similar taste to make predictions and recommendations. In this study, predictions are performed by matching user profile against the items available in the store. Thus, this study provides a recommendation for users by considering item to item match via user query in Business-to-Consumer (B2C) E-commerce.

INFORMATION STORAGE AND RETRIEVAL

Information storage and retrieval involves the process of finding those documents that are relevant to information need of the user from a given collection of documents (i.e., a complex information object, with components of different kinds such as text, images, video and sound). Radecki[4] present a new method of document retrieval based on the fundamental operations of the fuzzy set theory. Information storage and retrieval system simply refers to information retrieval system. Information Retrieval System (IRS) is a system that is capable of storage, retrieval and maintenance of information. Information in this context can be composed of text (including numeric and date), images, audio, video and other multimedia objects[5]. Although, the form of an object in an IRS is diverse, the text aspect has been the only data type that lent itself to full functional processing. The other data types have been treated as highly informative sources but are primarily linked for retrieval based upon search of the text.

An IRS consists of a software program that facilitates a user in finding the information the user needs. The system may use standard computer hardware or specialized hardware to support the search sub-function and to convert non-textual sources to a searchable media (e.g., transcription of audio to text). The gauge of success of an information system is how well it can minimize the overhead for a user to find the needed information. Overhead from a user's perspective is the time required to find the information needed, excluding the time for actually reading the relevant data. Thus search composition, search execution and reading non-relevant items are all aspects of information retrieval overhead. The IRSs originated with the need to organize information in central repositories (e.g., libraries) (Hyman-82). Catalogues were created to facilitate the identification and retrieval of items. Classical information retrieval models include the Boolean model which is simply based on set theory and queries as Boolean expressions, vector space model which queries and documents as vectors in term space and probabilistic model which adopts a probabilistic approach.

The general objective of an IRS is to minimize the overhead of a user locating needed information. Overhead can be expressed as the time a user spends in all of the steps leading to reading an item containing the needed information (e.g., query generation, query execution, scanning results of query to select items to read, reading non-relevant items). The success of an information system is very subjective, based upon what information is needed and the willingness of a user to accept overhead. Under some circumstances, needed information can be defined as all information that is in the system that relates to a user's need. In other cases it may be defined as sufficient information in the system to complete a task, allowing for missed data. For example, a financial advisor recommending a billion dollar purchase of another company needs to be sure that all relevant, significant information on the target company has been located and reviewed in writing the recommendation. In contrast, a student only requires sufficient references in a research paper to satisfy the expectations of the teacher which never is all inclusive. A system that supports reasonable retrieval requires fewer features than one which requires comprehensive retrieval. In many cases comprehensive retrieval is a negative feature because it overloads the user with more information than is needed. This makes it more difficult for the user to filter the relevant but non-useful information from the critical items.

Information retrieval, the term "relevant" item is used to represent an item containing the needed information. In reality the definition of relevance is not a binary classification but a continuous function. From a user's perspective, "relevant" and "needed" are synonymous. From a system perspective, information could be relevant to a search statement (i.e., matching the criteria of the search statement) even though it is not needed/relevant to user (e.g., the user already knew the information).

The two major measures commonly associated with information systems are precision and recall. Recall is the ratio of the number of relevant documents retrieved to the total number of relevant documents in the collection whereas precision is the ratio of the number of relevant documents retrieved to the total number of documents retrieved. When a user decides to issue a search looking for information on a topic, the total database is logically divided into four segments shown in Fig. 2.

Relevant items are those documents that contain information that helps the searcher in answering his question. Non-relevant items are those items that do not provide any directly useful information. There are two possibilities with respect to each item: it can be

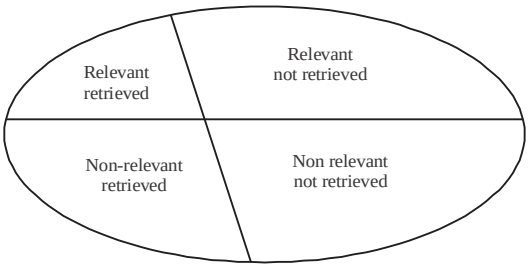

Fig. 2: Effects of search on total document space[6]

retrieved or not retrieved by theuser's query. Precision and recall are defined mathematically as in Eq. 1 and 2, respectively:

$$\text{Precision} = \frac{\text{Number\_retrieved\_relevant in answer}}{\text{Number\_total\_retrieved}} \quad (1)$$

$$\text{Recall} = \frac{\text{Number\_retrieved\_relevant in answer}}{\text{Number\_total\_relevant\_in collection}} \quad (2)$$

where, number_total_relevant_in collection is the number of relevant items in the database, number_total_retrieved is the total number of items retrieved from the query, number_retrieved_relevant in answer is the number of items retrieved that are relevant to the user's search need. Precision measures one aspect of information retrieval overhead for a user associated with a particular search. If a search has a 85% precision, then 15% of the user effort is overhead reviewing non-relevant items. Finally, high precision means few false alarms while high recall mean few false dismissals.

Indexing terms that are specific yields higher precision at the expense of recall. Indexing terms that are broad yields higher recall at the cost of precision. For this reason, an IR system's effectiveness is measured by the precision parameter at various recall levels. A single measure combining precision and recall, called F, expresses a compromise between precision and recall using harmonic mean method as follows:

$$F = \frac{2}{\frac{1}{r} + \frac{1}{p}} \quad (3)$$

In Eq. 3, r = recall, p = precision, F takes values in [0,1] such that $F \rightarrow 1$ as more retrieved documents are relevant and $F \rightarrow 0$ as few retrieved documents are relevant. Thus, F is high when both precision and recall are high.

Information retrieval models: An IR Model is characterized by four parameters:

- Representations for documents and queries
- Matching strategies for assessing the relevance of documents to a user query
- Methods for ranking query output
- Mechanisms for acquiring user-relevance feedback

IR Models can be classed into four types: set theoretic, algebraic, probabilistic and hybrid models. In the following sections, instances of each type is described in the context of the IR Model parameters.

Set theoretic models: The Boolean model represents documents by a set of index terms, each of which is viewed as a Boolean variable and valued as True if it is present in a document. No term weighting is allowed. Queries are specified as arbitrary Boolean expressions formed by linking terms through the standard logical operators: AND, OR, and NOT. Retrieval Status Value (RSV) is a measure of the query-document similarity. In the Boolean model, RSV equals 1 if the query expression evaluates to True; RSV is 0 otherwise. All documents whose RSV evaluates to 1 are considered relevant to the query.

This model is simple to implement and many commercial systems are based on it. User queries can employ arbitrarily complex expressions, but retrieval performance tends to be poor. It is not possible to rank the output, since, all retrieved documents have the same RSV, nor can weights be assigned to query terms. The results are often counter-intuitive. For example, if the user query specifies 10 terms linked by the logical connective AND, a document that has nine of these terms is not retrieved. User relevance feedback is often used in IR systems to improve retrieval effectiveness. Typically, a user is asked to indicate the relevance or irrelevance of a few documents placed at the top of the output. Since, the output is not ranked, however, the selection of documents for relevance feedback elicitation is difficult.

The fuzzy-set model is based on fuzzy-set theory, which allows partial membership in a set, as compared with conventional set theory which does not. It redefines logical operators appropriately to include partial set membership and processes user queries in a manner similar to the case of the Boolean model. Nevertheless, IR systems based on the fuzzy-set model have proved nearly as incapable of discriminating among the retrieved output as systems based on the Boolean model. The strict Boolean and fuzzy-set models are preferable to other models in terms of computational requirements which are low in terms of both the disk space required for storing document representations and the algorithmic complexity of indexing and computing query-document similarities.

Algebraic models: The vector-space model is based on the premise that documents in a collection can be

represented by a set of vectors in a space spanned by a set of normalized term vectors. If the vector space is spanned by n normalized term vectors, then each document will be represented by an n-dimensional vector. The value of the first component in this vector reflects the weight of the term in the document corresponding to the first dimension of the vector space and so forth. A user query is similarly represented by an n-dimensional vector. A query-document's RSV is given by the scalar product of the query and document vectors. The higher the RSV, the greater is the document's relevance to the query. The strength of this model lies in its simplicity. Relevance feedback can be easily incorporated into it. However, the rich expressiveness of query specification inherent in the Boolean model is sacrificed.

Probabilistic models: The vector-space model assumes that the term vectors spanning the space are orthogonal and that existing term relationships need not be taken into account. Furthermore, the model does not specify the query-document similarity which must be chosen somewhat arbitrarily.

The probabilistic model takes these term dependencies and relationships into account and, in fact, specifies major parameters such as the weights of the query terms and the form of the query document similarity. The model is based on two main para meters-Pr(rel) and Pr(nonrel), the probabilities of relevance and non relevance of a document to a user query which are computed using the probabilistic term weights and the actual terms present in the document.

Relevance is assumed to be a binary property so that Pr(rel) = 1-Pr(nonrel). In addition, the model uses two cost parameters, a1 and a2, to represent the loss associated with the retrieval of an irrelevant document and non retrieval of a relevant document, respectively.

The model requires term-occurrence probabilities in the relevant and irrelevant parts of the document collection which are difficult to estimate. However, this model serves an important function for characterizing retrieval processes and provides a theoretical justification for practices previously used on an empirical basis (for example, the introduction of certain term-weighting systems).

Hybrid models: As in the case of the vector-space model, the extended Boolean model represents a document as a vector in a space spanned by a set of orthonormal term vectors. However, the extended Boolean (or p-norm) model measures query-document similarity by using a generalized scalar product between the corresponding vectors in the document space. This generalization uses the well-known Lp-norm defined for an n-dimensional vector, d where the length of d is given by:

$$|d| = |(w_1, w_2, ..., w_n)| = \left(\sum_{j=1}^{n} w_j^p\right)\frac{1}{p} \quad (4)$$

In Eq. 4, $1 \le p \le \infty$ and $w_1, w_2, ..., w_n$ are the components of the vector d.

Generalized Boolean OR and AND operators are defined for the p-norm model. The interpretation of a query can be altered by using different values for p in computing query document similarity. When p = 1, the distinction between the Boolean operators AND and OR disappears as in the case of the vector-space model. When the query terms are all equally weighted and $p = \infty$, the interpretation of the query is the same as that in the fuzzy-set model. On the other hand, when the query terms are not weighted and $p = \infty$, the p-norm model behaves like the strict Boolean model. Varying the value of p from 1 to ∞ offers a retrieval model whose behavior corresponds to a point on the continuum spanning from the vector-space model to the fuzzy and strict Boolean models. The best value for p is determined empirically for a collection but is generally in the range $2 \le p \le 5$.

Query formulation and indexing: A query can be formulated by issuing an SQL command, or by proving an example document or image, or by browsing (e.g. displaying headers, summaries, miniatures for refining the retrieved results). The aim of a document retrieval system is to issue documents which contain the information needed by a given user of an information system[4]. The process of retrieving documents in response to a given query is carried out by means of the search patterns of these documents and the query. It is thus clear that the quality of this process, i.e., the pertinence of the information system response to the information need of a given user depends on the degree of accuracy in which documents and query contents are represented by their search patterns. Most document search patterns are sets of index-terms representing the ideas contained in the subject matter of the documents, sets of index-terms with numerical weights assigned to the terms according to their importance. While creating search patterns of documents and queries, one uses a thesaurus which is a set of terms on which specific kinds of relations are defined, e.g., the relation of synonymity, the relation of hierarchy, the relation of affinity. The weights assigned to index-terms and the use of a thesaurus during the creation of search patterns of documents and queries improve this process as well as the effectiveness of the document retrieval process. Search pattern of queries could be constructed using Boolean operators of AND, OR and NOT to connect the index-terms of queries. Indexing search documents that are likely to match the query.

Accuracy and speed are the goals of retrieval. The former retrieves documents that the user expects in the answer with as few incorrect answers (errors) as possible

and all relevant answers are retrieved. In the later, retrieval has to be fast as the system is expected to respond in real time. Accuracy of retrieval depends on what is matched with the query, the matching function, query criteria and complexity. The query is compared (matched) with all stored documents. The definition of similarity criteria is an important issue. Matching has to be fast. Also, document matching has to be computationally efficient and sequential searching must be avoided. The two types of errors are false dismissals or misses (qualifying but non retrieved documents) and false positives or false drops (retrieved but not qualifying documents). A good method minimizes both.

Similarity queries: Information retrieval is an approximate exercise where a query is specified and all documents up to a pre-specified degree of similarity are retrieved and presented to the user ordered by similarity. The two common types of similarity queries are the range queries which retrieve all documents up to a distance "threshold" T and the nearest-neighbor queries which retrieve the k best matches. Distance similarity decides whether two documents are similar such that the lower the distance, the more similar the documents are and also the higher the similarity, the more similar the documents are. For successful retrieval, the descriptions should be more accurate for a more accurate retrieval.

Text queries can be single or multiple keyword queries involving context queries (e.g., phrases, word proximity), Boolean queries (keywords with AND, OR, NOT operators), natural language (free text queries). Structured search takes also text structure into account and can be flat or hierarchical for searching in titles, paragraphs, sections, chapters or hypertext when combining content-connectivity. In keyword matching, the whole collection is searched and there is no preprocessing , no space overhead and updates are easy. The user specifies a string (regular expression) and the text is parsed using a finite state automaton. Typical algorithms used are KMP and BMH algorithms

Collaborative filtering: According to Francesco, etc., Collaborative Filtering (CF) is the process of filtering for information or patterns using techniques involving collaboration among multiple data sources, etc. In other words, it is a method of making automatic predictions about the interest of a user by collecting preferences or taste information from many users. The process of identifying similar users and recommending what similar users like is called collaborative filtering. Recommendation for a user U is then made by looking at the users that are most similar to U in this sense and recommending items that these users like.

Collaborative filtering explores techniques for matching people with interests and recommendations. It procedure often require user's active participation, an easy way to represent user's interests to the system and algorithms that are able to match people with similar interests. Typically, the workflow of a CF system is as follows:

- A user expresses his preferences by rating items (e.g., books, movies, etc.) of the system. These ratings can be viewed as an approximate representation of the user's interest in the corresponding domain
- The system matches this user's ratings against other user's and finds the people with most 'similar' tastes
- With similar users, the system recommends items that the similar users have rated highly but not yet being rated by this user (presumably the absence of rating is often considered as the unfamiliarity of an item)

Furthermore, a user's profile consists simply of the data the user has specified. This data is compared to those of other users to find overlaps in interests among users. These are then used to recommend new items. Essentially, each user has a set of nearest neighbors defined by using the correlation between past evaluations. Predicted scores for un-evaluated items of a target user are predicted by recommender system using a combination of the actual rating scores from the nearest neighbors of the target user. The goal of a collaborative filtering algorithm is to suggest new items or to predict the utility of a certain item for a particular user based on the user's previous likings and the opinions of other like-minded users. Figure 3 shows the schematic diagram of collaborative filtering process.

In a typical CF scenario, there is a list of m users = $\{U_1, U_2, U_3, \ldots, U_m\}$ and a list of n items = $\{i_1, i_2, i_3, \ldots, i_n\}$. Each user U_k has a list of items i_{Uk} which the user has expressed his opinions about. Opinions can be explicitly given by the user as a rating score, generally within a certain numerical scale or can be implicitly derived from purchased records by analyzing timing logs by mining web hyperlinks and so on[7]. There exist a distinguished user called the active user for whom the task of a collaborative filtering algorithm is to find an item likeliness that can be of two forms which are prediction and recommendation.

Prediction is a numerical value, P_{aj} expressing the predicted likeliness of item, for the active user. This predicted value is within the same scale (e.g., 1-5) as the opinion values provided by the active user. Recommendation is a list of N items that the active user will like most. The recommended list must be on items not already purchased by the active user. This interface of CF algorithm is also known as the Top_N recommendation.

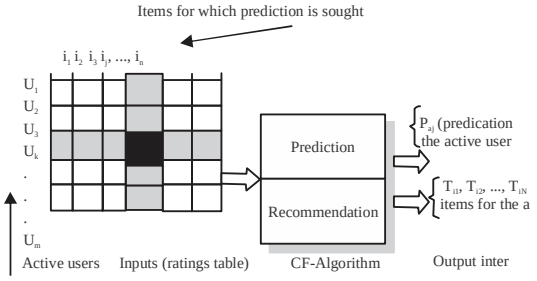

Fig. 3: The schematic diagram of collaborative filtering process[7]

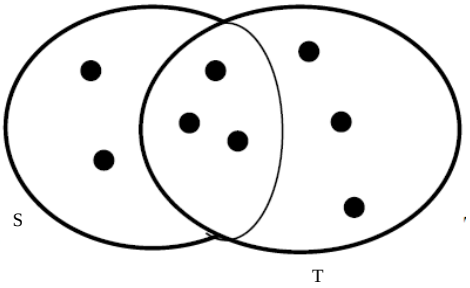

Fig. 4: Two sets with Jaccard similarity 3/8

Types of collaborative filtering: Collaborative filtering techniques can be classified into the various categories. These include memory-based, model based and hybrid techniques.

Memory-based collaborative filtering: The memory-based collaborative filtering uses user to user and item to item correlations based on rating behaviour to predict ratings and recommend items for the users in future. It is often referred to as neighborhood-based CF. This mechanism is used in many commercial systems as it is efficient and easy to implement.

Model-based collaborative filtering: Model-based CF algorithm uses recommender system information to create a model that generates the recommendations. Unlike memory-based CF, model based CF does not use the whole dataset to compute predictions for real data. There are various model-based algorithms including Bayesian networks, clustering models and latent semantic models such as singular value decomposition, principal component analysis and probabilistic matrix factorization for dimensionality reduction of rating matrix. The goal of this approach is to uncover latent factors that explain observed ratings[7].

Hybrid-based collaborative filtering: Hybrid CF algorithms are combinations of memory-based and model-based CF approaches which are used to overcome the drawbacks of memory-based and model based CF like sparsity and grey sheep. The essence is to improve the prediction performance of the CF algorithms.

Jaccard similarity of sets: A fundamental data-mining problem is to examine data for "similar" items. Similarity is one of the applications of near-neighbor search. One notion of similarity is the similarity of sets by looking at the relative size of their intersection. This notion of similarity is called "Jaccard similarity" and sometimes used in finding similar sets[6]. These include finding textually similar documents and collaborative filtering by finding similar customers and similar products. In a nut shell, Jaccard similarity is a statistic used for comparing the similarity and diversity of sample sets.

The Jaccard similarity of sets S and T is $|S\cap T|/|SUT| = S\cap T/|S|+|T|-|S\cap T|$, that is, the ratio of the size of the intersection of S and T to the size of their union. The Jaccard similarity of S and T is denoted by SIM(S, T) where 0 =SIM(S, T) = 1. If sets S and T are both empty, then SIM(S, T) = 1. Figure 4 represents two sets S and T with their intersection and union properties. There are three elements in their intersection and a total of eight elements that appear in S or T or both. Thus, SIM(S, T) = 3/8.

An important class of problems that Jaccard similarity addresses well is that of finding textually similar documents in a large corpus such as the Web or a collection of news articles. Here, the emphasis is on character-level similarity, not "similar meaning" which requires that one examines the words in the documents and their uses. Another class of applications where similarity of sets is very important is called collaborative filtering, a process whereby we recommend to users items that were liked by other users who have exhibited similar tastes. Amazon.com has millions of customers and sells millions of items. It database records which items have been bought by which customers. Two customers are said to be similar if their sets of purchased items have a high Jaccard similarity. Likewise, two items that have sets of purchasers with high Jaccard similarity will be deemed similar.

Modern information retrieval can be accessed from services of search engines such as Google, Yahoo, Bing, Excite, HotBot, Lycos, InfoSeek Guide, Alta Vista, Teoma, lxquick, Vivisimo, Don Busca, etc. The users can search for information in multimedia formats such as text, audio, still images and moving images[8] by looking up for keywords appeared in any documents and/or files stored in different formats such as HTML, MSWord, MSExcel, PDF and images. These documents and/or files which are distributed over large data source, will be stored on the

internet. As a result, in wide range information searching, searchers are not able to access the whole site causing incapability to obtain specific information.

Furthermore, searching results of meaning similarity and relation to keywords in some cases might not display required documents from specific keywords input. Thus, they are not able to find the document or web page they need. This can be as a result of lack of searching technique or knowledge of how to use specific keyword or keywords and search process. Keyword search is the simplest form of the most popular query method for search engine in information systems[8]. It contains a single keyword or multiple keywords and a short phrase. In a single keyword search, a particular word in the document will be displayed such as in a case of searching for sugar-producing crops. Keywords are specific words that can be sugarcane or we can query with the keyword in other forms to allow users to easily find the needed information quickly. The first significant issue that needs to be considered is the technique used to measure the similarity between a user specified key and the index finger to indicate directly to the required information.

Overview on search engines: The ability to search and retrieve information from the Web efficiently and effectively is an enabling technology for realizing its full potential. One way to find relevant documents on the Web is to launch a Web robot (also called a wanderer, worm, walker, spider or knowbot). These software programs receive a user query, then systematically explore the Web to locate documents, evaluate their relevance and return a rank-ordered list of documents to the user[3]. The vastness and exponential growth of the Web make this approach impractical for every user query. An alternative is to search a pre-compiled index built and updated periodically by Web robots. The index is a searchable archive that gives reference pointers to Web documents. This is obviously more practical and many existing search tools are based on this approach.

A search engine is an information retrieval system designed to help find information stored on a computer system. The search results are usually presented in a list and are called hits. Search engines help to minimize the time required to find information and the amount of information that must be consulted, thereby managing information overload. Web search engines have the advantage of offering access to a vast range of information resources located on the internet. Many search engines also search multimedia or other file types on the deep Web, often accessible as separate searches. Web search engines tend to be developed by private companies, though most are available free without charge. A Web search engine has three components viz, spider,

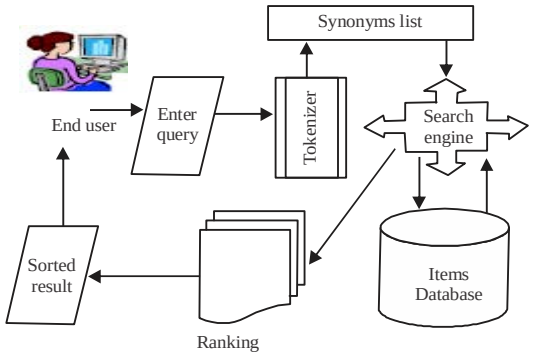

Fig. 5: Architecture of a typical search engine[9]

index and search engine mechanism. A spider is a program that traverses the Web link to link, identifying and reading pages. Spiders are said to 'crawl' the web in their hunt for pages to include in their search. The index is a database containing a copy of each Web page the spider gathers. The search engine mechanism refers to the software that enables users to query the index and results are usually returned in a ranked order. A more sophisticated development in search engine technology is the ordering of search results by concept, keyword, site, links or popularity. There is also the inclusion of artificial intelligence in determining what is relevant. Every search engine has rules for formulating queries. The spider finds pages and other contents in the pages, stores them in the database so that the database can be searched by keyword and whatever more advanced approaches adopted, and in the end the page will be found if the search matches its content. Figure 5 shows the architecture of a search engine. When a user types his query through the text field (provided for user's query entry), the tokenizer divides the user's query into words. The crawler or spider then locates the document's directories (e.g. Uniform Resource Locator, URL), keywords, metadata, and the title of document. The query of an end user that is divided into tokens is matched with the corresponding document terms, keywords, metadata, and the title in the database. The real documents whose keywords, metadata and title match the tokens of the end user's query are then retrieved to the document's directories.

Effective Web search is viewed as an information retrieval problem[1, 10]. IR problems are characterized by a collection of documents and a set of users who perform queries on the collection to find a particular subset of it. According to Gudivada *et al*.[3], comprehensively indexing the entire Web and building one huge integrated index will only further deteriorate retrieval effectiveness, since, the Web is growing at an exponential rate. On the other hand, a collection of Web indexes, each with its own specialized search tool, holds promise. Under this scheme, each Web index is targeted to comprehensively represent

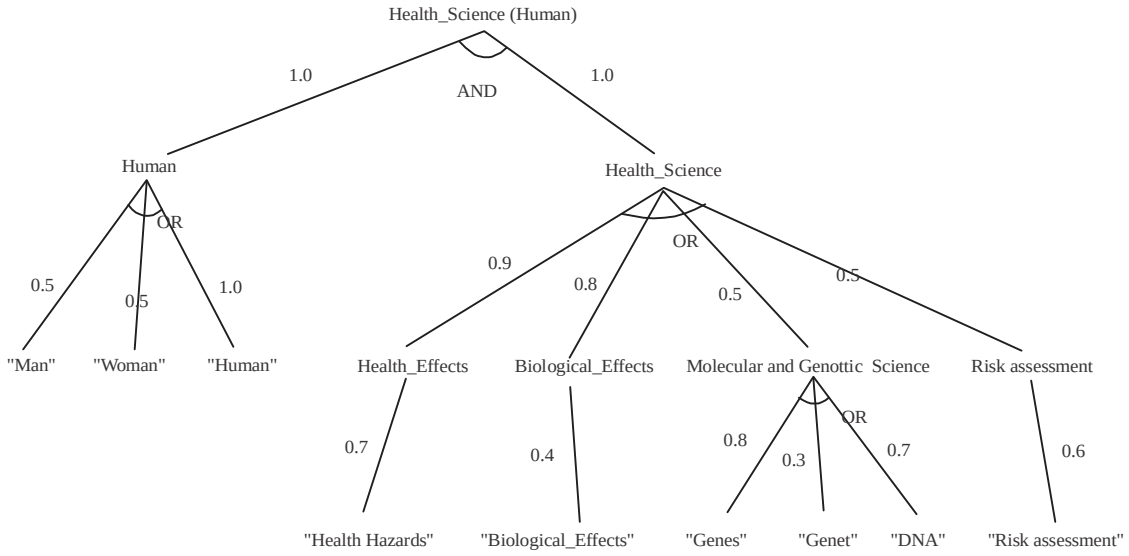

Fig. 6: Top-down approach of rule-based trees for concept-based retrieval

documents of a specific information space. Information spaces are bounded by for example, academic disciplines, a class of industries, a group of services. Most Internet search engines are based on the Boolean Retrieval Model. The model is relatively easy to implement. However, a few limitations of the model include:

- Inability to assign weights to query or document terms
- Inability to rank retrieved documents
- Naïve users have difficulty in using

Concept-based retrieval addresses the shortcomings of Boolean retrieval model by allowing search requests to be specified in terms of concepts structured as rule-base trees (Lu, etc.). The development of rule-base trees follows a top-down refinement strategy and enhances support for AND/OR relationships as well as support for user-defined weights as shown in Fig. 6.

E-commerce: E-commerce refers to the use of the Internet and the Web to transact business which involves the buying and selling of products and services. More formally, it is a digitally enabled commercial transaction between and among organizations and individuals, which seems to command the wave of the future. A typical E-commerce hub is sites like Amazon, ebaY, etc. Small business owners can benefit from E-commerce by selling their products on the Internet to customers that they were not able to reach before. This is called business-to-consumer E-commerce and involves customers gathering information, purchasing physical goods (i.e., books, clothes, phones, etc.) or information goods (i.e., digitized contents, movies, software, E-books, etc.). Small businesses can also benefit from E-commerce by buying products from larger companies to help run their business. This is called business-to-business E-commerce. However, there is also business-to-government E-commerce for public procurement, licensing procedures and other government related operations as well as consumer-to-consumer E-commerce between private individuals or consumers.

Various applications of E-commerce include online banking, online shopping and order tracking, domestic and international payment systems, document automation in supply chain and logistics, electronic tickets, digital wallets, instant messaging, social networking, and teleconferencing. One advantage of E-commerce is the ability to reach a global market without necessarily implying a large financial investment. The limits of this type of commerce are not defined geographically which allows consumers to make a global choice, obtain the necessary information and compare offers from all potential supplies, regardless of their locations. By allowing direct interaction with the final consumer, E-commerce shortens the product distribution chain, sometimes even eliminating it completely.

This way, a direct channel between the producer or service provider and the final user is created, enabling them to offer products and services that suit the individual preferences of the target market. However, issues of loss of privacy, cultural and economic identity, insecurity, lack of legislation to regulate its activities and strong dependence on information and communication technologies are problems.

LITERATURE REVIEW

The Tapestry system, Goldberge *et al*.[11] introduced collaborative filtering. In 1994, the GroupLens system [12] implemented a CF algorithm based on common user's preferences. Nowadays, it is known as memory-based CF algorithm because it employs users' similarities for the formation of the neighborhood of nearest users. Since then, many improvements of memory-based algorithms have been suggested [13].

According to Ekpenyong and Umoren[14], a Quality of Service (QoS)-Aware Signal to Interference and Noise Ratio (SINR)-based blocking probability model is simulated to evaluate the performance of networks with third generation (3G) interface.

In Umoren *et al*.[15], a computational intelligence framework for Length of Stay Prediction in Emergency Healthcare Services Department. the work carried out a computational intelligence framework based on fuzzy knowledge-based system and multiple criteria VHD algorithms is considered. The work proposed an Adaptive Intelligence Multi-Factored Algorithm (AIMFA) and multi criteria algorithm to predict and optimize handoff decision.

Karypis proposed another CF algorithm based on the items similarities for neighborhood generation. It is now referred to as item-based or item-by-item CF algorithm because it employs item's similarities for the formation of neighborhood of nearest users. Most recent work followed the two aforementioned approaches (i.e., user-based or item-based). Herlocker *et al*.[13] weigh similarities by the number of common ratings between users/items, when it is less than some common threshold parameter γ. Deshpande and Karypis[16] applied item-based CF algorithm combined with conditional-based probability similarity and Cosine similarity measures. Xue *et al*.[17] suggest a hybrid integration of aforementioned algorithms (nearest neighbor CF algorithms) with model-based CF algorithms. Finally, recent extensions of CF include issues like streaming data[18] or privacy preserving[19]. This work shall design and develop an information retrieval system for item-based collaborative filtering using Jaccard similarity algorithm.

SYSTEM DESIGN

This is the process of defining the architecture, components, modules, interfaces and data for a system to satisfy specified requirements. The design of a system comes after the analysis and the goal is to produce a model or representation of an entity that will be built. The software design approach adopted in this work is the object-oriented analysis and design methodology and the use of unified modeling language (UML) diagrams for illustrations.

System description: Collaborative filtering using Jaccard similarity is a model used to filter an item in order to present a relevant list to the user. A user starts by entering search term and the system calculates Jaccard similarity between the search term and all the items in the database irrespective of the item's category. Based on the results obtained (values >0.0 are accepted), an item will be retrieved and displayed as a searched result. When a user selects an item to view and/or buy, the system increments the item's rank by 0.5 and additional items are suggested to the user based on the rank of all items that are related to the search item.

The justification for the proposed system will enhance competitive advantage since a company that has high reliability index for its product will have an advantage over those with low reliability index. It will improve customer's loyalty by recommending a list of items that could customers could buy in addition to certain goods they purchase. Furthermore, the company's reputation will increase with increase customers' satisfaction leading to improved patronage on their part and more profit generation on the part of the company.

System architecture: System architecture refers to a formal description and representation of a system organized in a way that supports reasoning about the structures and behaviors of the system. The architecture of the proposed collaborative filtering using Jaccard similarity is presented in Fig. 7.

Existing users have their data (such as user name, password, customer name, customer address, gender, phone number, etc.) already stored in the database for user profiling. When an active user arrives the system as a new user, his/her data has to be logged in and stored in the database equally using authenticated username and password. However, an active user whether existing or new can search for and view an item through a query by typing the item name. The tokens in the name are then sent to the collaborative filtering system to find out which category list the item belongs. There is a process of synchronization with the database and the information is sent to the Jaccard system where product ranking is performed by the similarity co-efficient algorithm. Based on the results obtained, a list of items and their prices will be recommended to the active user for purchase to be made. This recommended list is very useful as it reminds the user what to purchase along with selected item based on his or her preferences. Finally, the objective of an IRS which is to minimize the overhead of a user locating needed information is achieved.

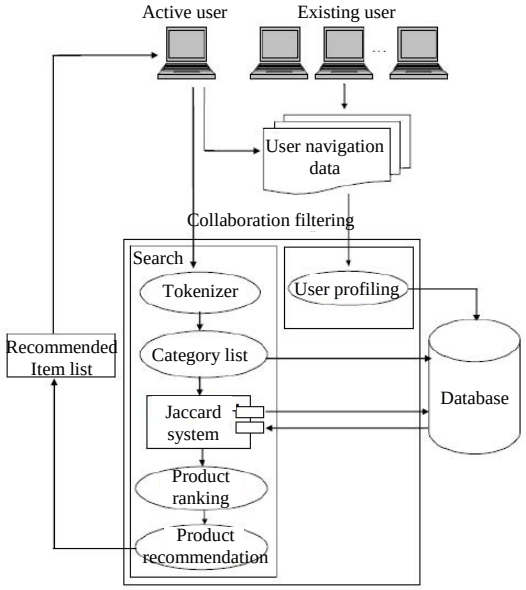

Fig. 7: Proposed system architecture for memory-based collaborative filtering

Table 1: User table

Filename	Data types	Size	Description
C_name	Varchar	35	Customer name
C_address	Varchar	30	Customer address
Gender	Number	6	Gender
Email	Varchar	20	Customer email address
Phone	Varchar	11	Phone number
N_bus_stop	Varchar	25	Nearest bus stop
Street	Varchar	25	Street name
Username	Varchar	15	Username
Password	Varchar	8	Password

Table 2: Item table

Filename	Data types	Size	Description
Item_name	Varchar	35	Item name
Item_category	Varchar	30	Item category
Item_quantity	Number	30	Item quantity
Item_description	Varchar	3	Item description
Item_price	Currency	10	Item price
Item_rank	Number	5	Item rank

Database design: This is the definition of the database tables used in the Jaccard similarity based collaborative filtering system. These tables include user table, item table and purchase table (Table 1-3). The user table on Table 1 shows the store customer's profile that will be used during product delivery.

The item table store all the items used by the collaborative filtering system and the schema is shown on Table 2.

The purchase table stores information related to the purchase of an item by a customer. Such information includes customer name and address, item name, item quantity, item price and item category. The schema is shown on Table 3.

Table 3: Purchase table

Filename	Data types	Size	Description
C_name	Varchar	35	Customer name
C_address	Varchar	30	Customer address
Item_name	Varchar	35	Item name
Item_category	Varchar	30	Item category
Item_quantity	Number	30	Item quantity
Item_price	Currency	10	Item price

Proposed system algorithm: The proposed system uses Jaccard similarity coefficient algorithm developed by Paul Jaccard to calculate the similarity between two sets. This algorithm is given in Eq. 5:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (5)$$

For example, given two sets A and B as follows:

- A = {'Techno', Nokia', 'Phone', 'Samsung', 'Toy', 'Laptop'}
- B = {'Samsung, 'Apple', 'Phone', 'Laptop'}

Then, $|A \cap B|$ = {'Phone', 'Samsung', 'Laptop'} and $|A \cup B|$ = {'Techno', 'Nokia', 'Phone', 'Samsung', 'Toy', 'Laptop', 'Apple'} and J(A, B) = 3/7 = 0.43.

Algorithm 1; The Jaccard intersection algorithm:

```
start
initialize set A[ ]
initialize set B[ ]
initialize counter = 0
        for i = 0 to A. length
                for j = 0 to B.length
                    If (A [i] = = B[j])
                    counter++;end for
         end for
return counter; end
```

Algorithm 12; The Jaccard Union Algorithm:

```
start
initialize set A[ ]
initialize set B[ ]
initialize counter = 0
initialize set C = A
        for i = 0 to B. length
                initializeadd_variable = false
        for j = 0 to A. length
                if (B[i] = = A[j])
                add_variable = true
        break
                end if
        end for
       if (add-variable)
       C.add(A[i])
       end if
      end for
      return C. length
stop
```

Use case design: A use case diagram consists of actors, use cases and their relationships. Figure 8 is used to model the system by capturing the functionality of the system. The use case diagram for the propose system is shown in Fig. 8.

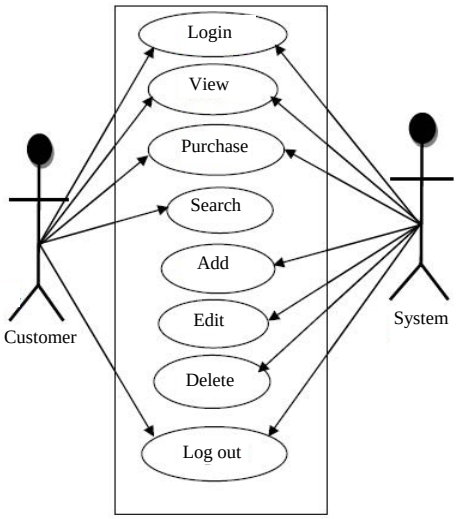

Fig. 8: Use case diagram for B2C E-commerce

Activity diagram: The activity diagram describes the dynamic aspect of the system. it is basically a flow diagram to represent the flow from one activity to another in the operation of the system. It is shown in Fig. 9.

Class diagram: This diagram represents the static view of the system and is used for visualizing, describing and documenting different aspects of the system. It is also used for constructing the executable codes of the proposed collaborative filtering system for items retrieval. The class diagram is as shown in Fig. 10.

INPUT/OUTPUT DESIGN

Input design: The input design to the system is the Jaccard Search form which accepts a user's query and carries out Jaccard Similarity computation so as to retrieve relevant items to be displayed to the user. This form is presented in Fig. 11.

Output design: The output of this system is the items displayed to the user based on the search term. The Jaccard search form is used in displaying these items and is shown in Fig. 12.

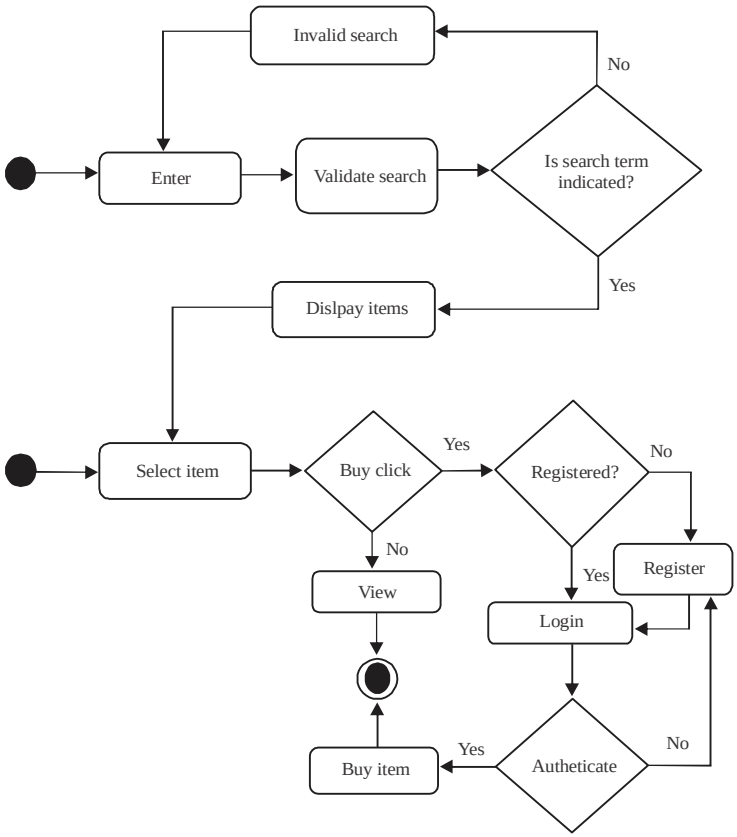

Fig. 9: Activity diagram of the proposed system

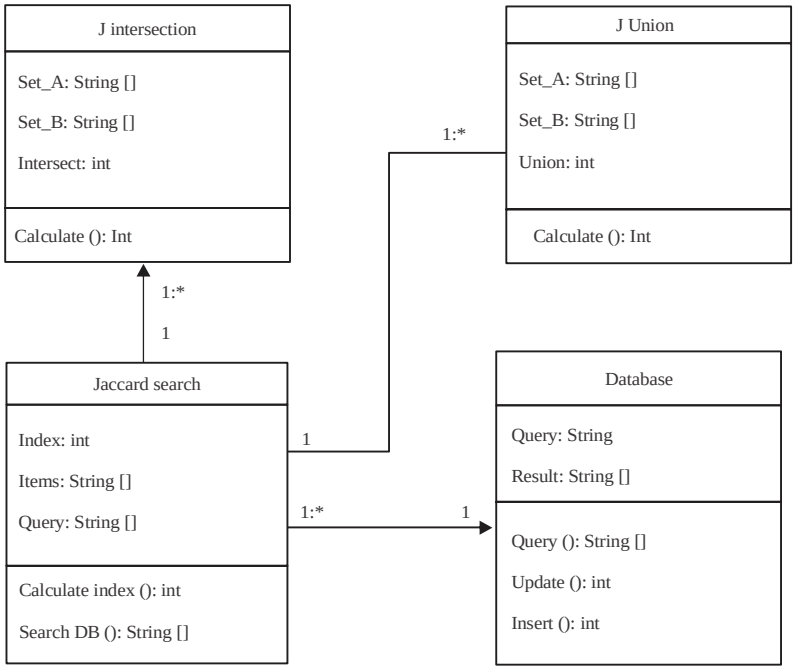

Fig. 10: Class diagram of the proposed system

Fig. 11: Input design

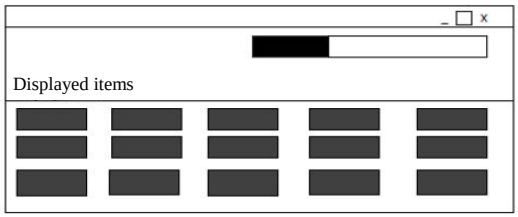

Fig. 12: Output design

SCREEN SHOTS OF THE APPLICATION

The following graphical user interface provides a means of interaction between the user and the developed system.

Homepage: The Homepage in Fig. 13 welcomes the user to the online store and introduces the user to the system. The online user can decide to search for items in the store, or select from the items displayed on the welcome page. The page is made up of categories of items available in the store as well as randomly generated items displayed for the user's view.

Search page: The search page in Fig. 14 returns results from entered query. The user is required to type in the desired search term after which the "Search" button is clicked to display its related results in a results area component. On-clicking the search button, the result is sorted in order of similarity by enabling the "Jaccard Similarity" with a user's threshold value of zero (0).

Items view page: This page shown in Fig. 15 displays details of a particular item selected.

Purchase page: This log in page in Fig. 16 grants the user access to purchase the selected it emby logging in as an existing user or purchase can be made by registering as a new user as shown in Fig. 17.

Database of items: Figure 18 shows the items in the database and their descriptions. The item terms are queried and presented to the user in the search result page. These queries are carried out by the server in the back end using a scripting language called PHP. New items can also be added into the system by the administrator using the "Insert" statement in PHP. The database was designed and developed in MySQL.

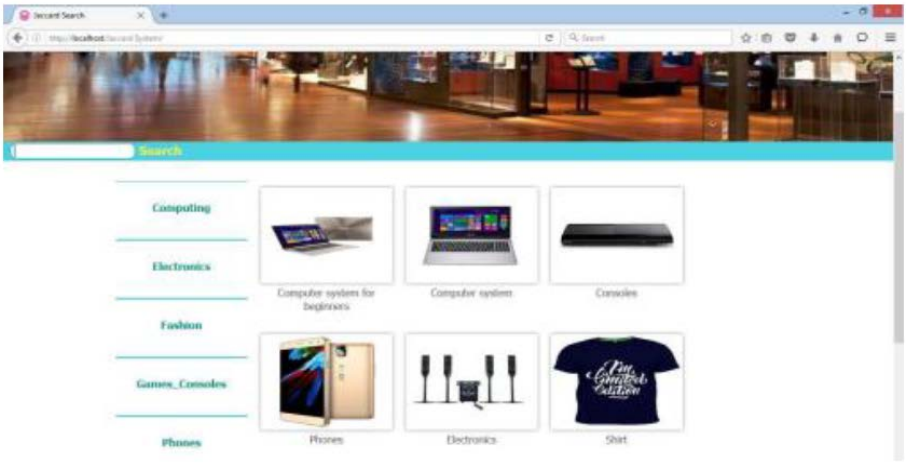

Fig. 13: Homepage of the system

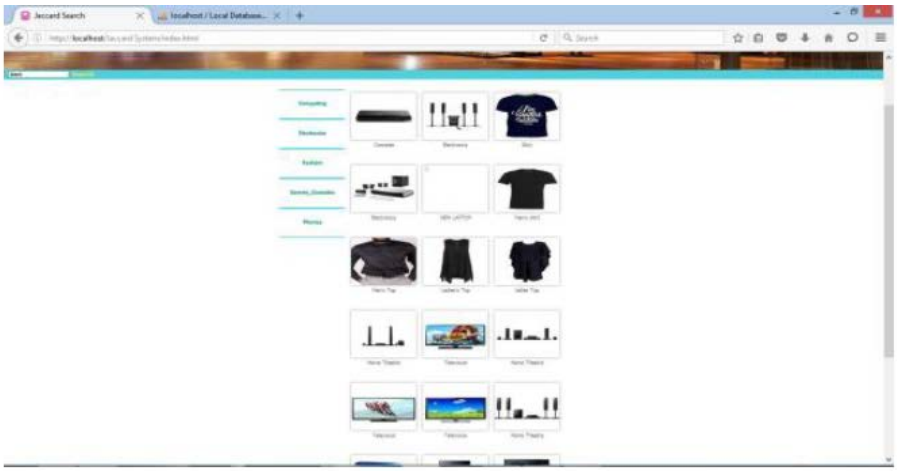

Fig. 14: Search results page

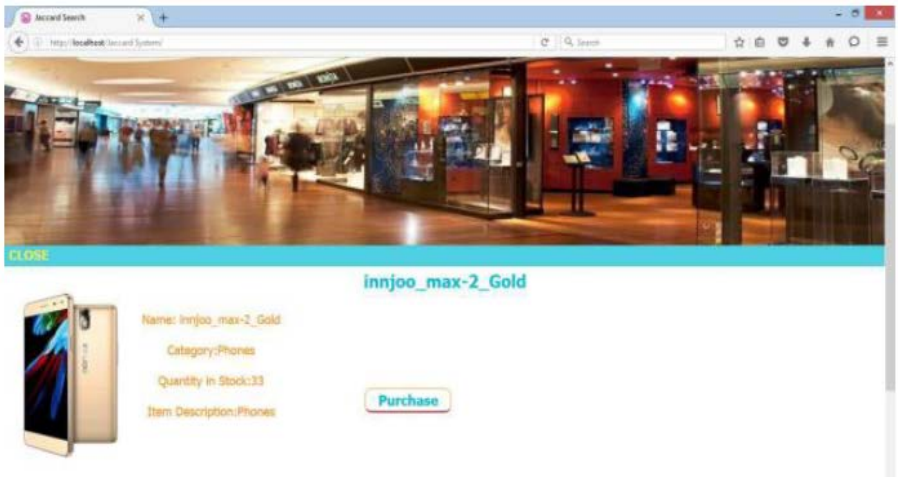

Fig. 15: Database of items

Fig. 16: Purchase log in page

Fig. 17: Customer's registration page

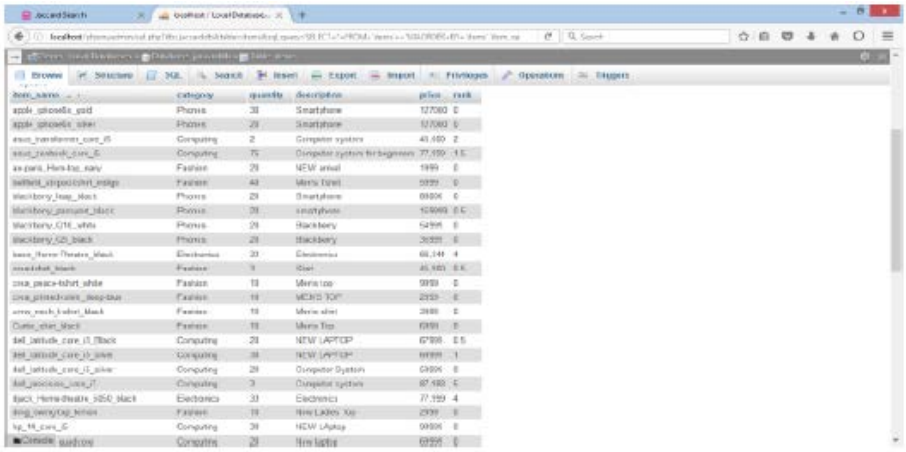

Fig. 18: Database of items

EVALUATION OF SEARCH RESULTS

From the search result in Fig. 14, the following assertions were made:

- Total Items in store = 50
- Total number of items retrieved = 20
- Total number of total relevant = 30
- Number relevant retrieved (Phones) = 4

$$\text{Precision} = \frac{\text{Number}_{\text{of relevant retrieved}}}{\text{Total}_{\text{item retrieved}}} \times 100\% =$$

$$\frac{4}{20} \times 100\% = 20\%$$

$$\text{Recall} = \frac{\text{Relevant}_{\text{retrieved}}}{\text{Total}_{\text{relevant}}} \times 100\% =$$

$$\frac{4}{30} \times 100\% = 13.33\%$$

CONCLUSION

The developed memory-based collaborative filtering system is efficient and user-friendly with the use of a search algorithm. The combination of Java and HTML was useful in the design of a user-friendly interface for the web application while all the items were stored in MySQL database system. The system could store large amount of data without compromising its speed and accuracy of retrieval. The results obtained from the system indicates that it provides easy retrieval of items to meet personalized and recommended lists and so is useful for small or medium-scale enterprise that are interested in managing their customer profiles for loyalty and revenue generation.

RECOMMENDATIONS

The developed system is recommended for small and medium-scale enterprise to help create awareness for online users, enhance business growth and maximize profit for the organization. New items can be easily stored irrespective of size and retrieval is easy with the use of Jaccard similarity algorithm. The following recommendations are made for smooth operation, use and maintenance of the system:

- Awareness should be created to indeed users in order for them to start using the system
- Adequate security control through antivirus software and intrusion detection mechanism should be provided to ensure smooth running of online operations
- All installations and operations procedures must be adhered to, so as to derive maximum benefit from using the system

REFERENCES

01. Salton, G., 1989. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley Publishing Co., MA.
02. Faloutsos, C., 1996. Searching Multimedia Databases by Content. Kluwer Academic Publishers, New York, USA.,.
03. Gudivada, V.N., V.V. Raghavan, W.I. Grosky and R. Kasanagottu, 1997. Information retrieval on the world wide web. IEEE Internet Comput., 1: 58-68.
04. Radecki, T., 1976. Application of fuzzy set theory to the description of information retrieval process. Communications of the Main Library and Scientific Information Centre of the Technical University of Wroclaw, Poland.
05. Kowalski, G.J. and M.T. Maybury, 2002. Information Storage and Retrieval Systems: Theory and Implementation. 2nd Edn., Kluwer Academic Publishers, New York, USA.,.
06. Rajaraman, A., J. Leskovec and J.D. Ullman, 2012. Mining of Massive Datasets. Stanford University Press, California, USA.,.
07. Sarwar, B., G. Karypis, J. Konstan and J. Reidl, 2001. Item-based collaborative filtering recommendation algorithms. Proceedings of the 10th International Conference on World Wide Web, May 1-5, 2001, Hong Kong, China, pp: 285-295.
08. Supachai, T., 2010. System for Storage and Retrieval of Information by Computer. Pithak Printing Press, Bangkok, Thailand,.
09. Akinribido, C.T., B.S. Afolabi, B.I. Akhigbe and I.J. Udo, 2011. A fuzzy-ontology based information retrieval system for relevant feedback. Int. J. Comput. Sci., 8: 382-389.
10. Frakes, W.B. and R. Baeza-Yates, 1992. Information Retrieval Data Structures and Algorithms. Prentice-Hall Inc., Englewood Cliffs, New Jercy,.
11. Goldberg, K., T. Roeder, D. Gupta and C. Perkins, 2001. Eigentaste: A constant time collaborative filtering algorithm. Inform. Retrieval J., 4: 133-151.
12. Resnick, P., N. Lakovou, M. Sushak, P. Bergstrom and J. Riedl, 1994. Group lens: An open architecture for collaborative filtering of Netnews. Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, Oct. 22-26, Chapel Hill, North Carolina, United States, ACM Press, pp: 175-186.

13. Herlocker, J.L., J.A. Konstan, A. Borchers and J. Riedl, 1999. An algorithmic framework for performing collaborative filtering. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, California, pp: 230-237.
14. Ekpenyong, M. and I. Umoren, 2012. QoS-aware SINR-based call blocking evaluation in cellular networks with 3G interface. Int. J. Comput. Sci. Issues (IJCSI.), 9: 441-446.
15. Umoren, I., K. Udonyah and E. Isong, 2019. A computational intelligence framework for length of stay prediction in emergency healthcare services department. Proceedings of the 2019 14th International Conference on Computer Science & Education (ICCSE), August 19-21, 2019, IEEE, Toronto, Canada, pp: 539-549.
16. Deshpande, M. and G. Karypis, 2004. Item-based top-n recommendation algorithms. ACM Trans. Inform. Syst., 22: 143-177.
17. Xue, G.R., C. Lin, Q. Yang, W. Xi and H.J. Zeng *et al*., 2005. Scalable collaborative filtering using cluster-based smoothing. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 2005, ACM, Salvador, Brazil, ISBN:1-59593-034-5, pp: 114-121.
18. Barajas, J. and X. Li, 2005. Collaborative filtering on data streams. Proceedings of 9th European Conference on Principles and Practice of Knowledge Discovery on Databases, October 3-7, 2005, Springer, Porto, Portugal, pp: 429-436.
19. Polat, H. and W. Du, 2005. Privacy-preserving collaborative filtering on vertically partitioned data. Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery, October 3-7., 2005, Springer, Berlin, Germany, pp: 651-658.