# Optimal page ranking system for web page personalisation using kernel-based fuzzy C-means and gravitational search algorithm

P. Pranitha*

University of JNTUH,
Hyderabad, Telangana, India
Email: pranithap0784@gmail.com
*Corresponding author

M.A.H. Farquad

Faculty of Computers and Information Systems,
Islamic University of Madinah, Saudi Arabia
Email: farquadonline@gmail.com

G. Narshimha

JNTUH,
Karimnagar, Telangana, India
Email: narsimha06@gmail.com

**Abstract:** In this personalised web search (PWS), we utilise a kernel-based FCM for clustering a web pages. For effective personalised web search, queries are optimised using GSA with respect to clustered query sessions. In offline processing, initially preprocess the input information taken from consumer visited web pages and are transformed in to numerical matrix. These matrices are gathered with the help of kernel-based FCM method after produce a vector for consumer query and detect a minimum distance as centroid values these values are input to the GSA algorithm. It will engender these links given top N web pages from cluster. In online processing, the user query is engaged as input then extract some web pages from Google, Bing, Yahoo also extract content and snippet from web pages. Finally, detect a sum of contents and snippets and web pages would be considered in descending order.

**Keywords:** kernel-based fuzzy c-means; clustering; offline; online; preprocessing; Google; Bing; Yahoo.

**Reference** to this paper should be made as follows: Pranitha, P., Farquad, M.A.H. and Narshimha, G. (2020) 'Optimal page ranking system for web page personalisation using kernel-based fuzzy C-means and gravitational search algorithm', *Int. J. Business Intelligence and Data Mining*, Vol. 16, No. 1, pp.1–19.

**Biographical notes:** P. Pranitha obtained her BTech in Computer Science Engineering from University of JNTUH, Hyderabad, Telangana, India. Then she obtained her MTech in Software Engineering from University of JNTUH, Hyderabad, Telangana, India and pursuing her PhD in Computer Science Engineering majoring in Web Mining from University of University of

JNTUH, Hyderabad, Telangana, India. Currently, she is working as an Associate Professor in the Department of Computer Science and Engineering, JITs, Karimnagar, University of JNTUH, Hyderabad, Telangana, India. Her specialisations include data mining, web mining, social networks, classification, link mining and clustering. Her current research interests are optimal page ranking system for web page personalisation.

M.A.H. Farquad is currently working as an Assistant Professor in the Faculty of Computers and Information Systems, Islamic University of Madinah. His research interests include supervised learning (support vector machine, neural network, decision tree, etc.) and unsupervised learning (k-means). His research mainly focuses on application of machine learning algorithm in various fields viz. bankruptcy prediction, insurance fraud detection, churn prediction and function approximation problems, etc. He published his research work in esteemed journals, i.e., *Information Systems*, *Expert Systems with Applications* and *Decision Support Systems*. He also engaged in reviewing activities of various Elsevier and IEEE journals and conferences.

G. Narshimha is working as an Associate Professor at JNTUH, Karimnagar, Telangana, India. He has completed his BE in ECE at Osmaniya University, Hyderabad and obtained Master in CS&E in 1999 at Osmaniya University. He has awarded doctorate in CS&E Osmaniya University Hyderbad, India in July 2009. He has about 17 years of teaching experience. He has published 70 papers in both national and international conferences followed by 38 international and national journals. His interested areas are computer networks, mobile computing, network security, cloud computing and data mining. Five PhD are awarded and 13 research scholars are working under him. He is life member of Indian society for technical education.

## 1 Introduction

The fast development in electronics and internet technologies has meaningfully impacted society and our daily life. At the time of the past few years, the World Wide Web has become the biggest and the most popular manner of communication and data dissemination. Every day, the web grows by roughly millions of electronic pages, adding to the hundreds of millions of pages already online (Castellano et al., 2011). Web-based technologies for computers and also for mobile devices were utilised to give a user friendly and easy interaction within our search engine and the end users (Dao et al., 2013). This web is extensively utilised by individuals and organisations in numerous fields, like e-banking, education, e-commerce, research, news distribution, entertainment, and communication (Wen et al., 2012). Even though it is utilised in numerous implementations it has been pointed out that the occur of a large number of smart things on the web causes great problems to computers and also humans to detect, pick, and use smart things in an operative manner. Because of the issue of data overload, it is enormously difficult for web users to progress the situation awareness about the huge number of things introduced on the web everyday (Cai et al., 2014). Further as the number of web pages has exploded, it has become significant to establish them in order to restrict the search space for data (Jeong et al., 2014).

To facilitate this web page recommendation is utilised. Web page recommendation has become progressively popular, and is displayed as links to associated stories,

associated books, or most viewed pages at websites. When a user browses a website, a system of visited web pages during a term (the period from starting, to existing the browser by the user) can be produced (Nguyen et al., 2014). The exemplary search solutions ranking performed by web search engines can be a beneficial model for other data systems providers, especially libraries, to imitate. Since people are now used to web search interfaces and relevancy-ranked results lists, they imagine searching in library catalogues to be as informal, and the presentation of solutions to be as good, as when they search the web (Behnert and Lewandowski, 2015). Henceforth web page ranking is one among the significant aspects and is very much needed to attain the relevant pages on the basis of the interest of the user from the large assortment of disordered data. Search engines are a significant tool for users to recover data for a specific query. The notion of building a web search engine was predominantly appealing as it would involve the interoperability of dissimilar technologies, like crawling the Semantic Web with the aid of intelligent agents (Batzios and Mitkas, 2012). The web search engine has long become the most significant portal for ordinary people looking for useful data on the web. A search task begins with a query that the user problems to a search engine. The search engine procedures the query and returns a search solutions page showing an ordered list of sites that may encompass the data the user wants (Wang et al., 2015). The user scrutinises the excerpts of the sites, picks a few sites, and browses the nominated documents. Users may experience failure when search engines yield in appropriate results that do not meet their real and anticipated intentions. Such inappropriate think is largely because of the huge variety of users' contexts and backgrounds, and also the ambiguity of the texts (Malthankar and Kolte, 2016).

Henceforth, current search engines cannot fully please the user's requirement for high-quality data search services; and it raises numerous novel challenges for data retrieval (Hariharan et al., 2015). Personalised searches that have appeared in response to such tests, produce users' search results depended partly on their personal interests and/or past search histories; thus, personalised searches can display dissimilar search results for dissimilar users to make each search more pertinent to users' needs (Kim et al., 2012). User-centred design (UCD) methods can be utilised to augment the personalisation capabilities of web-based schemes with adaptive navigation support via recommendations (Santos et al., 2014). There are dissimilar obtainable approaches to grade the recommended action in the guidelines, each focusing on assured query criteria like quality of the evidence, benefits, harms, generalisability and applicability, and also patient preferences, ethical, political and economic factors (Khodambashi et al., 2015). The strength of this united query assortments can substantially augment the utilisation of query suggestion to progress web search quality (Li et al., 2011).

## 2 Literature survey

With the debauched expansion of information innovation, the present time is seeing an exponential augmentation in the era and gathering of web data. Forestalling the right information to the opportune individual is turning out to be more troublesome stage by stage that thus adds multifaceted nature to the elementary leadership prepare. Recommendation outlines are shrewd outlines that address this issue. They are normally used as a portion of web-based business sites to indorse items to clients. Mainstream of

the recommendation proposal deliberate just the substance information of clients and disregard sequential information. Sequential data similarly provides cooperative experiences about the conduct of clients. Mishra et al. (2015) have built up a new outline that considered sequential information display in web route designs, alongside substance information. They additionally measured delicate clusters amid clustering that aided in catching the numerous interests of clients. The projected outline had utilised comparability upper assessment and singular value decomposition (SVD) for the epoch of proposals for clients. They tried their method on three datasets, the MSNBC benchmark dataset, and replicated dataset and CTI dataset. They compared their method and the chief request Markov display and also random expectation show. The outcomes permitted the reasonability of their approach.

The occurrence of web search engines (WSEs) authorises them to generate a substantial measure of data in the kind of question logs. These records comprise of all search inquiries put composed by clients. Practical compensations could be earned by technique of offering or discharging those logs to strangers. Though, this data possibly uncovers sensitive client information. Emptying direct identifiers is not satisfactory to save the privacy of the clients. Some current privacy-preserving methodologies use log bunch preparing at the similar time, as logs are produced and devoured in an ongoing domain, a ceaseless anonymisation technique would be more obliging. Estrems et al. (2016) projected:

1 another approach to anonymise inquiry logs, in the aspect of k anonymity

2 some de-anonymisation apparatus to choose conceivable privacy problems, on the off chance that that an aggressor retrieved the anonymised question logs.

That technique saved the first client interests, though spread conceivable semi-identifier information over many clients, anticipating linkage assaults. To survey that is execution, all the projected algorithms were objectified and a broad preparation of analyses was led utilising genuine data.

Chawla (2016) transported a method for producing the optimal ranked clicked URLs using genetic algorithm (GA) in light of clustered web question sessions for powerful personalised web search (PWS). Exploratory study was led on the dataset of web inquiry sessions caught in the spaces scholastics, amusement and games to test the adequacy of cluster wise ideal ranked clicked URLs for PWS. The outcomes that were confirmed measurably, established a change in the normal exactness of the PWS in light of ideal positioned clicked URLs over both classic IR and PWS without ideal positioned clicked URLs. In that way, the viability of PWS using ideal positioned clicked URLs was confirmed for better personalising the web search as per the information requirement of the client.

Web search engines are revolving into a noteworthy phase for the general population to get to information. It has been suggested that because of the search patterns of search engines clients are resembled with rising occasions, the question log of search engines has the capacity for trend surveillance. For instance, observing flare-ups of scourges. Numerous trend surveillance examines have discovered the utilisation of inquiry logs and have strived to discriminate question terms reasonable for trend surveillance. The mainstream of this work selects delegate question terms by counselling space specialists or by setting up a substantial content corpus for feature extraction. The process of these methodologies, in any case, is too exclusive to make the trend surveillance strategies

versatile to numerous themes. Fang and Chen (2015) familiarised a versatile trend surveillance strategy. They built up a straightforward and convincing feature extraction algorithm, called TF-LTR that influenced the record returned through search engines and the frequency of the terms in the refunded archives to select agent inquiry terms of slanting points. In specific, they scrutinised combine insightful figuring out to rank models keeping in mind the end goal to gauge a term's discriminative power in creating a report rank greater in the returned record list.

A great portion of the current web search preparations are worked down for fulfilling wide preparation of clients irrespective of innocent or experts. Then, with the rise of fast web implementations and propelled Web 2.0-based rich internet applications (i.e., online journals, wikis, and so on.), it has turned out to be much simpler for clients to allocate data over the web. This brings a restraint for web search answers for let singular clients locate the right information rendering to their predispositions. In this article, we portray our method of authorising personalised web scan for clients in light of their predispositions. It is a restraint in itself to have the predispositions of the clients known to and considered through search engines. Shafiq et al. (2015) have self-possessed and built up their remarkable method of perceiving the dispositions of clients from the important portions of their interpersonal organisations and groups. They trusted that the data recognised with the inquiries postured by clients might have solid relationship with important information in their interpersonal organisations. With a precise end goal to discover the individual interests and the social-settings, they exposed:

1 exercises of clients in their social-networks

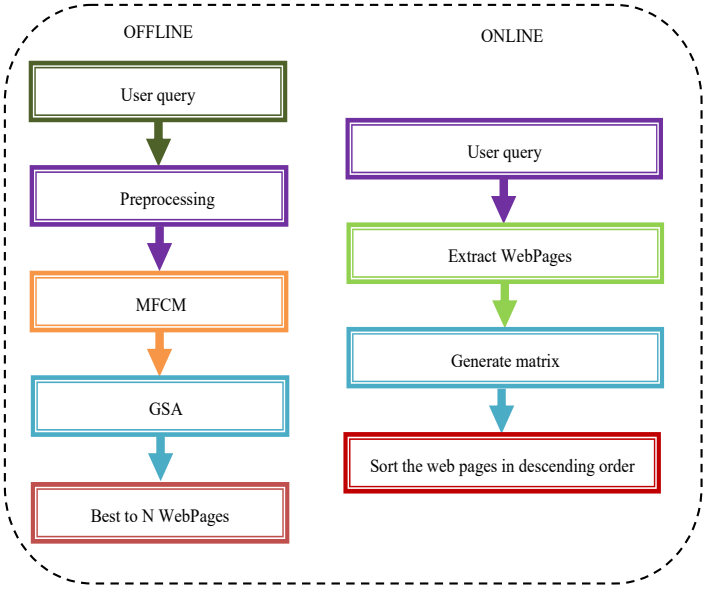
2 significant information from client's social-networks, in view of their projected trust and implication lattices.

Bouadjenek et al. (2016) inspected a commitment towards the amalgamation of social information in the ordering structure of an IR outline. Since every client had his/her own specific comprehension and viewpoint of a provided archive, they projected a method in which the file design gave a personalised social document representation (PerSaDoR) of every report per client in light of his/her exercises in a social labelling outline. The projected method depended on lattice factorisation to register the PerSaDoR of records that coordinated a question, at inquiry time. The intricacy examination established that their method scaled straightly with the quantity of records that matched the inquiry, and in that manner, that could scale to expansive datasets. PerSaDoR had been similarly seriously evaluated by a disconnected study and by a client study worked on a substantial open dataset from scrumptious indicating critical compensations for personalised search compared with best in class policies.

Xu et al. (2014) focused on the problem of creating temporal semantic context for ideas. The objective of the projected problem was to elucidate a notion with temporal, succinct, and organised data that could mirror the express and faceted insinuations of the idea. The temporal semantic context could aid clients learn and comprehend new or recently rose ideas. The projected temporal semantic context structure combined the rudiments from word reference, Wikipedia, and LinkedIn sites. An overall plan to create temporal semantic context of a notion by building that is linked words, related ideas, setting sentences, setting diagram, and setting groups was projected. Exact trials on three characteristic datasets including Q-A dataset, LinkedIn dataset, and Wikipedia dataset established that the projected algorithm was effectual and precise. Diverse from

physically twisted context archives, for instance, LinkedIn and Wikipedia, the projected method could obviously create the context and did not necessitate any earlier learning, for instance, cosmology or numerous levelled database.

**Figure 1** Proposed optimal page ranking system for web page personalisation (see online version for colours)




## 3 Problem definition

In this segment discuss the problem definition of my research work,

- Current recommendation system shows confident limitations like intelligence, adaptability, flexibility, limited accuracy.
- Because of non-existence of accuracy, extended and high run time available recommendation schemes exhibit the issues of less coverage.
- Pages that are newly added or rarely visited by end user are not displayed by the available method that also a significant problem.
- Designing a recommendation scheme that contemplates sequential data is still a significant problem.

- The statistical clustering only gives the crisp clustering; that does not match with the real world implementations.
- In available web page recommendation scheme, the users considered only the sequential aspect of a web user session with the help of sequential mining algorithms which provides only the patterns that exist in the systems. In our work, we have proposed a system that produces the recommendations to the users, in view of the sequential data that exist in their usage designs of web pages.

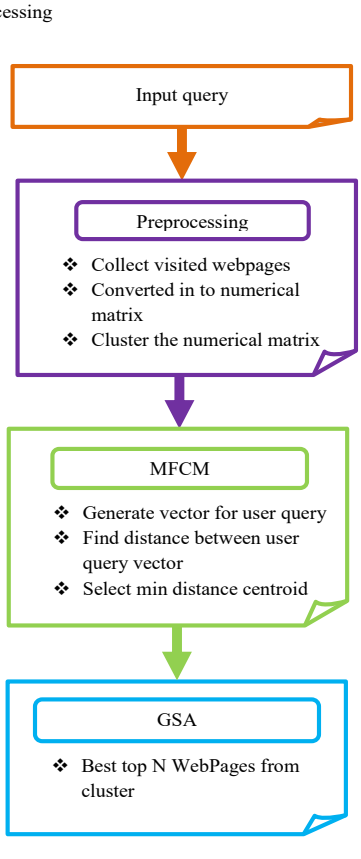
## 4 Proposed methodology

The main intension of this research is to personalise the web search with the help of optimal ranked URLs that overcomes the disadvantages of links analysis and PWS ranking. The proposed methodology encompasses of two phases offline and online. In the offline processing, the query sessions on web are administered to produce the query session keyword vectors. The query session keyword vectors will be clustered in groups with the help of modified kernel-based fuzzy c-means algorithm (MKFCM). The optimisation algorithm will be utilised on each clusters to detect the optimal subset of clicked pages that are not only relevant measured with the help of high Information Scent but also have satisfactory internal dissimilarity in order to have extensive coverage of sub-domains characterised in the cluster. Gravitational search algorithm (GSA) is utilised for detecting the optimal clusters in that the content similarity will be measured as the fitness function. So that, we can overcome spam pages in top ranked documents, at the time of the personalisation of web search, the optimal ranked clicked URLs related to selected cluster are recommended to the user on each requested result page. At the time of online processing, the user input query is utilised to pick the most similar cluster measured with the help of semantic-based similarity measure. Once the cluster is selected, the sub clusters of a selected cluster are searched to detect the sub cluster that is most similar to the data need of the input query. If the selected sub cluster is more comparable to the input query than the whole cluster, then high scent web pages related to the sub cluster are recommended otherwise the optimal subset of pages related to the cluster0 covering the dissimilar sub-domains of the cluster is recommended. Thus in this technique of personalisation of web search it is the selection of cluster that adapts according to the user profile not the optimal ranking of the clicked URLs that has already been computed in offline processing. The application will be done in JAVA and the function of the proposed technique will be analysed with numerous available system.

### *4.1 Offline processing*

In the offline processing, primarily the input data's are composed from user-visited web pages then these web pages are converted in to matrix format. At the time of offline processing the modified FCM is implemented on query session clustered using in order to recognise the subset of pages related to a given cluster. Thus at the end of offline processing, each cluster is related to optimal subset of web pages. It will cluster a numerical matrix in to dissimilar set.

**Figure 2** Working process of offline (see online version for colours)




### *4.1.1 Modified kernel-based fuzzy C-means*

In this page ranking-based web page personalisation at this time we will cluster a query keyword session vector from these query we will cluster with the help of modified kernel base FCM the query's are clustered in to form a group of clusters. In numerous kernel fuzzy c-means algorithmic procedure enlarges the kernel fuzzy c-means algorithm with a dissimilar kernel learning setting. The proposed technique use multiple kernel fuzzy c means for clustering the executed jobs. The objective performance of proposed multiple kernel fuzzy c means algorithm is efficiently elucidated as follows.

$$F(M, A) = \sum_{i=1}^{N} \sum_{j=1}^{a} M_{ij}^{m} \left(1 - K_{MK}\left(t_j, A_i\right)\right) \tag{1}$$

where

$M_{ij}$ is the membership of $j^{th}$ data in the $i^{th}$ cluster $A_i$

$A$ is the cluster centre

$K_{MK}$ is the multiple kernel function.

#### *4.1.1.1 Procedure for MKFCM*

Step 1 Initialise the number of task ($t$), number of cluster ($A$) and number of kernels ($K$).

Step 2 Initialise the membership matrix $M$.

Step 3 Compute the cluster centre by the subsequent equation

$$A_j = \frac{\sum_{i=1}^{N} M_{ij}^{m} t_i}{\sum_{i=1}^{N} M_{ij}} \tag{2}$$

Step 4 Update the membership function by the subsequent equation

$$M_{ij} = \frac{1}{\sum_{K=1}^{A} \left( \frac{\| t_i - A_j \|}{\| t_i - A_K \|} \right)^{\frac{2}{m-1}}} \tag{3}$$

$m$ is any real number greater than 'one'.

In multiple kernel fuzzy c means, $t_i$ signifies the kernel function $k_{MK}(x, y)$. At this time, we are considering multiple kernels for our projected work. So $k_{MK}(x, y) = k1(x, y) + k2(x, y)$ is a kernel.

$$K_{MK}(x, y) = K_1(x, y) + K_2(x, y) \tag{4}$$

$$K_1(x, y) = x^T y + c \tag{5}$$

$$K2(x, y) = 1 - \frac{\| x - y \|^2}{\| x - y \|^2 + c} \tag{6}$$

where $c$ is the constant value.

From the above equation the cluster centre equation (4) and membership equation (5) is adapted. Now the cluster centre calculation is performed with the help of equation (6),

$$A_j = \frac{\sum_{i=1}^{N} M_{ij}^{m} k_{MK}(x, y)}{\sum_{i=1}^{N} M_{ij}} \tag{7}$$

Membership updation is done by equation (7),

$$M_{ij} = \frac{1}{\sum_{K=1}^{A} \left( \frac{\left\| K_{MK}(x, y) - A_j \right\|}{\left\| K_{MK}(x, y) - A_K \right\|} \right)^{\frac{2}{m-1}}} \tag{8}$$

Step 5 if $\| M^{(K+1)} - M^{(K)} < \in$ then stop, otherwise go to equation (8).

On the basis of the MKFCM the executed jobs or tasks are clustered. At this time, we will cluster a numerical matrix into dissimilar clusters.

### *4.1.2 Query session keyword vector*

In our projected web page personalisation system, initially we will alteration an input query as query session keyword for this procedure we will use this equation 1. Now, the query session of URL is taken form user clicked URL in the search engine. The solution of the input query was submitting in a query and it designates only those session is measured. In Each query session keyword vector is produced from query session that is characterised as follows,

$$\text{Query Session} = (\text{Input Query},\ (\text{Click URLs/Page})^{+}) \tag{9}$$

where clicked URLs are those URLs that user clicked in the search solution of the input query before submitting another query; '+' specify only those sessions are considered that have at least one clicked page related to the input query. The query session vector $Q_i$ of the $i^{th}$ session is well defined as linear combination of contents vector of each clicked page $P_{id}$ scaled by the weight Sid that is the data scent related to the clicked page $P_{id}$ in session *I*.

$$Q_i = \sum_{d=1}^{n} S_{id} * p_{id} \qquad \forall i \in 1 \ldots m \tag{10}$$

In equation (9) n is the number of distinct clicked pages in the session *i* and $s_{id}$ (information scent) is considered for each clicked page present in a provided session *i* as defined in equation (9). The content vector of clicked page $P_{id}$ each $i^{th}$ query session is attained as weighted vector $Q_i$ with the help of formula (9). This vector is modelling the information need associated with the $i^{th}$ query session with the assistance of these we will generate a vector for user query also detect a distance between user query vector with cluster centroid from these centroid we will detect a minimum distance centroid. Lastly these minimum distance clusters are given in to GSA.

### *4.1.3 Gravitational search algorithm*

GSA is stimulated by the law of gravity and the law of motion. The algorithm is gathered under population-based method consisting of different masses. The masses share information to direct the search towards the best location in the search space, based on the gravitational force. GSA is based on the Newton's law of gravity and the Newton's law of motion as shown as equations (11) and (12).

$$F_{ij}^{d}(t) = G_c(t)\frac{M_{ai}(t)\times M_{pj}(t)}{D_{ij}^{2}} \tag{11}$$

$$A_i(t) = \frac{F_{ij}^{d}(t)}{M_i(t)} \tag{12}$$

where

$M_{ai}$ active gravitational mass associated to the agent $i$

$M_{pj}$ passive gravitational mass linked to $j$

$G_c(t)$ gravitational constant

$D_{ij}$ Euclidian distance value

$A_i(t)$ acceleration

$M_i(t)$ inertial mass.

The agents are supposed as objects and their function appraised by their masses in this algorithm. With the gravity force all objects attract each other that activate a global crusade of all the objects to those with the heavier masses. Hereafter, the masses work together via a direct form of communication via the gravitational force. The slow movement of the heavier masses settles the development phase of the method and is associated with an outstanding result. In the GSA, each mass (agent) comprehends four specifications such as the position, inertial mass, active gravitational mass, and the passive gravitational mass. The position of the mass indicates a result of the problem, and its gravitational and inertial masses are definite by a fitness function. Consequently, each mass bids a result, and the method is directed by properly adapting the gravitational and inertia masses. The masses get attracted with the help of the heaviest mass that intensely brings an optimum result in the search space. In this gravitational algorithm is utilised to find top N web pages from a cluster.

Step 1 Initialisation of agents

The position of the $N$ number of agents is arbitrarily initialised and the population is characterised as the available resource; the position of the $i^{\text{th}}$ agent is represented with the help of the subsequent relation:

$$R_i = \left(R_i^1, R_i^2, \ldots, R_i^d, \ldots, R_i^n\right) \quad for = (i = 1, 2, \ldots, N) \tag{13}$$

where

$R_i^d$ resource position of $i^{\text{th}}$ agent in the $d^{\text{th}}$ dimension.

Step 2 Fitness calculation

The fitness evolution is performed with the help of assessing the best (*bst*) and worst (*wst*) fitness for all agents at number of iteration, for minimisation issues.

$$Fit(t) = \min \cos t \tag{14}$$

$$bst(t) = \min_{j \in (1,..,N)} Fit_j(t) \tag{15}$$

$$wst(t) = \max_{j \in (1,...N)} Fit_j(t) \tag{16}$$

Step 3 Gravitational constant ($G$) computation

The gravitational constant $G$ is set at the beginning and is scaled down over a period of time in order to attain the search precision. Consequently, $G$ indicates a function of the initial value $G_{initial}$ and time ($t$);

$$G_c(t) = G\left(G_{initial}, t\right) \tag{17}$$

By means of fitness assessment, we competently compute the gravitational and inertia masses. A heavier mass displays an enormously efficient agent. In other words, the outstanding agents are endowed with greater attractions and walk further slowly. Presuming the equality of the gravitational and inertia mass, the values of masses is evaluated by the map of fitness.

Step 4 Masses of the agent calculation

Gravitational and inertia masses for each agent are intended at iteration $t$,

$$M_{ai} = M_{pj} = M_i$$

where $i = 1, 2, \ldots, N$

$$m_i(t) = \frac{Fit_i(t) - wst(t)}{bst(t) - wst(t)} \tag{18}$$

$$M_i(t) = \frac{m_i(t)}{\sum_{j=1}^{N} m_j(t)} \tag{19}$$

where

$Fit_i(t)$ signifies the fitness value of the agent $i$ at time $t$.

Step 5 Accelerations of agent calculation

Then the acceleration of agent $i$ at time $t$ can be articulated as:

$$A_i(t) = \frac{F_{ij}^d(t)}{M_i(t)} \tag{20}$$

Step 6 Velocity and positions of agents

Furthermore, the subsequent velocity of an agent is deemed as a fraction of its current velocity added to its acceleration. Henceforth, its position and its velocity are assessed by means of the equations given below.

$$V_i^d(t+1) = rand_i \times V_i^d(t) + A_i(t) \tag{21}$$

$$R_i^d(t+1) = R_i^d(t) + v_i^d(t+1) \tag{22}$$

where

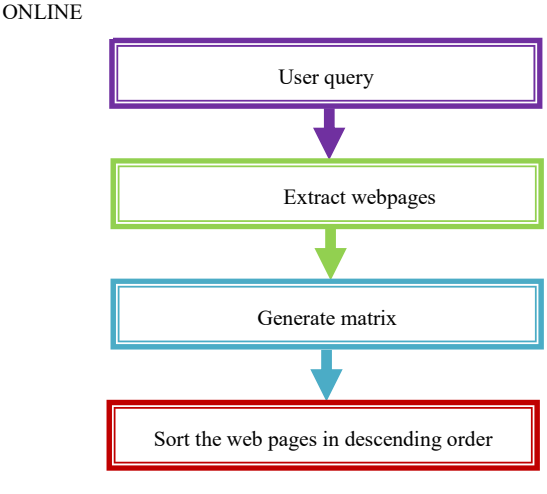
$rand_i$ uniform random variable in the interval [0, 1].

The optimisation algorithm will be utilised on each cluster to detect the optimal subset of clicked pages. The input for this GSA algorithm is minimum distance of clusters from these we will detect a best top N web pages from the cluster.

## *4.2 Online processing*

In the online processing in the input user query the web pages are occupied from some search engines like Google, Yahoo, Bing these are utilised for to gather a user query from these we will extract a content and snippet from each web pages.

**Figure 3** Working process of online (see online version for colours)




### *4.2.1 Snippet-based calculation*

The snippet is a small piece of data about the chief data in the link. The snippet would be visible under each link we desisted from the search engine as a small note. The calculation on the basis of snippet of the unique link is as follows: we patterned the query

word and the meanings of the query word with the snippet of each link to compute the number of occurrence in the snippet.

$$S_s(p) = \sum_{i=1}^{a} \left( \frac{S_s D_i}{\max(SD_i)} \times w_Q + \sum_{j=1}^{b} \frac{S_s D_i M_j}{\max(SD_i M_j)} \times w_M \right) \tag{23}$$

In the above equation, $S_s(p)$ is the considered snippet-based value of $s^{th}$ unique link; and $S_sD_i$ is the number of incidence of $i^{th}$ query word $D$ in the snippet $S$ of $s^{th}$ unique link; and $\max(SD_i)$ is the maximum number occurrence of $i^{th}$ query word $D$ in the snippet $S$ of whole unique links we attained; and $S_sD_iM_j$ is $j^{th}$ meaning of $i^{th}$ query word $D$ in the snippet $S$ of $s^{th}$ unique link; and $\max(SD_iM_j)$ is the maximum number of occurrence of $j^{th}$ meaning of $i^{th}$ query word $D$ in the snippet $S$ of the whole unique links we attained; and $w_Q$ is the weight value of the query word; and $w_M$ is the weight value of the meaning of the query word.

### *4.2.2 Content-based calculation*

In content-based calculation, we associate the contents of each link with the disconnected query words and their synonyms to check the number of incidence of disconnected query words and their synonyms in the contents of each link.

$$C_s(p) = \sum_{i=1}^{a} \left( \frac{C_s D_i}{\max(CD_i)} \times w_Q + \sum_{j=1}^{b} \frac{C_s D_i M_j}{\max(CD_i M_j)} \times w_M a \right) \tag{24}$$

In the above equation, $C_s(p)$ is the intended content-based value of $s^{th}$ unique link; and $C_sD_i$ is the number of occurrence of $i^{th}$ query word $D$ in the content of $s^{th}$ unique link; and $\max(CD_i)$ is the maximum number of occurrence of $i^{th}$ query word $D$ in the content of $s^{th}$ unique link; and $C_sD_iM_j$ is the number of occurrence of $j^{th}$ synonym of $i^{th}$ query word $D$ in the content of $s^{th}$ unique link; and $\max(CD_iM_j)$ is the maximum number of occurrence of $j^{th}$ synonym of $i^{th}$ query word $D$ in the content of $s^{th}$ unique link. Lastly produce a trust matrix for both content and snippet-based matrix on the basis of these we will detect the sum of both trust matrix and sum of each row it would be taken in descending order to sort the web pages.

## 5 Result and discussion

This segment shows the result we attained for our proposed technique in comparison with the existing technique. Our technique is implemented in Java (jdk 1.7) that has system configuration of Core 2 duo processor with clock speed of 2.3 GHZ and RAM of 2 GB and that runs Windows 7 OS.

### *5.1 Query description and our process*

This section explains the queries we used for our comparison. We took this input from webpage visited by the user from each search engine and the search engines we used are ‘Google’, ‘Bing’ and ‘Yahoo’. Now, we have ten links from user searched links collected from different search engine. To compare the response time between the proposed

technique and the existing technique, the experimentation is done using ten links from each search engine.

**Table 1** Optimal page ranking description and web page personalisation

| *Rank* | *Links* | *G* | *B* | *Y* |
|---|---|---|---|---|
| 1 | https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=0ahUKEwjYn6ywx_jQAhVMRY8KHdoWBKwQFggmMAE&url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FData_mining&usg=AFQjCNGDF7f8edqCNhLP4NS_x3dRKgBVxw | 1 | 1 | 1 |
| 2 | http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm | 2 | 2 | 2 |
| 3 | https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=5&cad=rja&uact=8&ved=0ahUKEwjYn6ywx_jQAhVMRY8KHdoWBKwQFgg_MAQ&url=http%3A%2F%2Fwww.thearling.com%2Ftext%2Fdmwhite%2Fdmwhite.htm&usg=AFQjCNFxnDUF4uNQxo0_7T8jiapPthy9-w | 4 | 10 | 4 |
| 4 | https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=12&cad=rja&uact=8&ved=0ahUKEwjYn6ywx_jQAhVMRY8KHdoWBKwQFghpMAs&url=http%3A%2F%2Fwww.laits.utexas.edu%2F~anorman%2FBUS.FOR%2Fcourse.mat%2FAlex%2F&usg=AFQjCNGgQxGvNXbd3WnVw9YhyJZWd98kkQ | 7 | 11 | 8 |
| 5 | http://www.tutorialspoint.com/data_mining/ | 9 | 11 | 11 |
| 6 | https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=8&cad=rja&uact=8&ved=0ahUKEwjYn6ywx_jQAhVMRY8KHdoWBKwQFghRMAc&url=http%3A%2F%2Fwww.sas.com%2Fen_us%2Finsights%2Fanalytics%2Fdata-mining.html&usg=AFQjCNGGjFqar9oc-wyZqh5HcRPDJmZUGQ | 5 | 5 | 11 |
| 7 | http://searchsqlserver.techtarget.com/definition/data-mining | 3 | 3 | 11 |
| 8 | https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=10&cad=rja&uact=8&ved=0ahUKEwjYn6ywx_jQAhVMRY8KHdoWBKwQFghcMAk&url=http%3A%2F%2Fwww.businessdictionary.com%2Fdefinition%2Fdata-mining.html&usg=AFQjCNGcw0wb4NCFX3xp0VH80TqLihL3Ew | 6 | 4 | 3 |
| 9 | https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=11&cad=rja&uact=8&ved=0ahUKEwjYn6ywx_jQAhVMRY8KHdoWBKwQFghiMAo&url=http%3A%2F%2Fwww.webopedia.com%2FTERM%2FD%2Fdata_mining.html&usg=AFQjCNEXMJBd99ewbwqwoeEVNaMoUeDJVQ | 11 | 9 | 7 |
| 10 | https://www.coursera.org/specializations/data-mining | 11 | 11 | 9 |

Table 1 explains as follows: the first column represents the ranking for the query second column represents a user searched links taken from different search engines we have taken ten search links from each search engine. Third column represents a Google search engine fourth column represents a Bing search engine fifth column represents a Yahoo search engine.

### 5.2 Performance comparison

This section shows the performance of our technique compared to the existing technique and individual web search engines such as 'Google', 'Bing' and 'Yahoo'. The performance is calculated based on the precision. The precision is calculated for the queries given by the user. The precision is calculated by taking the total relevant documents retrieved for the query divided by total documents retrieved for the query.

#### 5.2.1 Precision based on user given queries

The precision using user given queries is explained in this section. Figure 4 shows the precision comparison for the user given query when top ten links are taken from each search engine.

**Figure 4** Precision comparison for the user search link 'www.anderson.ucla.edu' (see online version for colours)

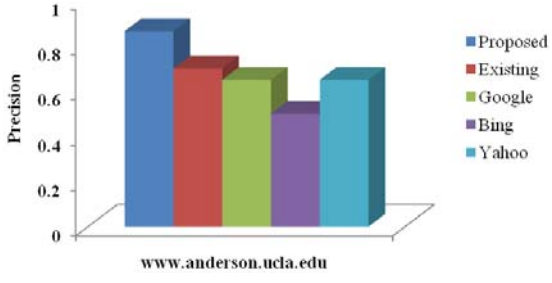



Figure 4 shows the precision of our technique compared to the existing FCM technique with our proposed KFCM the search engines we used for our technique for the user given first three links are taken. Here, the precision of our KFCM technique is high compared to the other FCM techniques. The precision we obtained for our technique is 76% and the precision obtained for the existing technique is 70% and the precision obtained using Google and Yahoo is 60% and the precision obtained using Bing is 50%.

**Figure 5** Precision comparison for the user search link 'www.threaling.com' (see online version for colours)

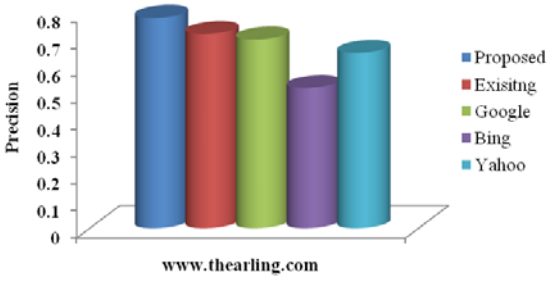


Figure 5 shows the precision of our technique compared to the existing technique with the search engines we used. Here, the precision of our KFCM technique is high compared to the other existing FCM techniques. The precision we obtained for our technique is 76% and it is 72% using existing technique and it is 70% using Google search engine and it is 50% using Bing search engine and it is 60% using Yahoo search engine.

**Figure 6** Precision comparison for the user searched link 'www.laits.utexas.edu' (see online version for colours)

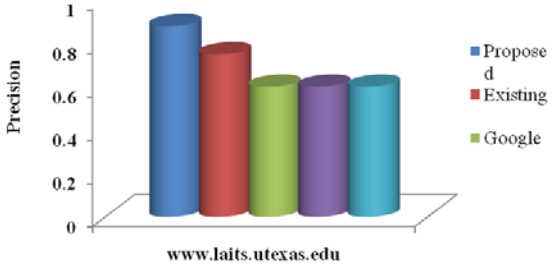
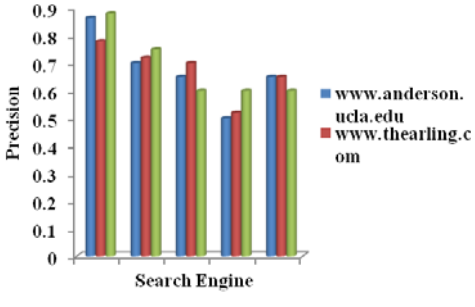



Figure 6 shows the precision comparison of our KFCM technique compared to the existing FCM technique for the user query 'www.laits.utexas.edu' when top ten links are considered. Here, the precision of our technique is high compared to the existing FCM technique. Figure 6 shows the precision comparison for the user given queries user searched links are taken from each search engine.

**Figure 7** Precision comparison for the user given queries taken from search engine (see online version for colours)




In Figure 7, the precision is calculated for each user given query by considering top hundred links from each search engine. The comparison shows that the proposed technique achieved better precision than the existing technique for all the user given queries.

### 5.2.2 Response time

Table 2 shows the response time comparison of our technique with the existing FCM technique in terms of top ten links. After giving the query, top ten links from each search engine are taken and merged based on the unique links. Thereafter the queries are ranked using our KFCM technique and the existing technique. The values in Table 2 is in milliseconds and the first column shows the queries we used for our comparison and N represents the list which derives from the fusion of the input rankings and tpr represents the time taken to rank the merged list. In most cases, the time taken to rank the merged list of our technique is less compared to the existing FCM technique when top ten links are taken. Table 2 shows the response time comparison in terms of top hundred links.

**Table 2** Response time comparison in terms of user searched links

| *Query* | *N* | *tpr (in ms)* | |
|---|---|---|---|
| | | *Our technique* | *Old technique* |
| Data mining | 10 | 2,364 | 3,154 |
| Image processing | 10 | 3,254 | 4,265 |
| Software engineering | 10 | 2,965 | 3,268 |
| Data mining and image processing | 10 | 2,648 | 4,685 |

## 6 Conclusions

In this paper, we proposed kernel-based fuzzy c-means clustering technique and GSA. In our proposed methodology mainly consider two main process such as offline and online processing. In offline processing initially preprocess a data the input data taken from user visited web pages and these web pages are converted in to numerical matrix these matrix are clustered with the aid of kernel-based FCM technique after generate a vector for user query and find a minimum distance as centroid values these values are input to the GSA algorithm. It will generated these links a given a top N web pages from cluster. In online processing the user query is taken as input then extract some web pages from Google, Bing, Yahoo also extract content and snippet from web pages. Finally, the find a sum of contents and snippets and web pages would be taken in descending order. In most cases, the response time based on top 50, ten links of our technique is better compared to the existing FCM technique and the precision of our kernel-based FCM technique is high compared to the existing technique.

## References

Batzios, A. and Mitkas, P.A. (2012) 'WebOWL: a semantic web search engine development experiment', *Elsevier on Expert Systems with Applications*, Vol. 39, No. 5, pp.5052–5060.

Behnert, C. and Lewandowski, D. (2015) 'Ranking search results in library information systems – considering ranking approaches adapted from web search engines', *Elsevier on the Journal of Academic Librarianship*, Vol. 41, No. 6, pp.725–735.

Bouadjenek, M.B., Hacid, H., Bouzeghoub, M.B. and Vakali, A. (2016) 'PerSaDoR: personalized social document representation for improving web search', *Elsevier on Information Sciences*, pp.1–35.

Cai, Y., Lau, R.Y., Liao, S.S., Li, C., Leung, H.F. and Ma, L.C. (2014) 'Object typicality for effective web of things recommendations', *Elsevier on Decision Support Systems*, Vol. 63, pp.52–63.

Castellano, G., Fanelli, A.M. and Torsello, M.A. (2011) 'NEWER: a system for neuro-fuzzy web recommendation', *Elsevier on Applied Soft Computing*, Vol. 11, No. 1, pp.793–806.

Chawla, S. (2016) 'A novel approach of cluster based optimal ranking of clicked URLs using genetic algorithm for effective personalized web search', *Elsevier on Applied Soft Computing*, Vol. 46, pp.90–103.

Dao, T.T., Hoang, T.N., Ta, X.H. and Tho, M.C.H.B. (2013) 'Knowledge-based personalized search engine for the web-based human musculoskeletal system resources (HMSR) in biomechanics', *Elsevier on Journal of Biomedical Informatics*, Vol. 46, No. 1, pp.160–173.

Estrems, D.P., Roca, J.C. and Viejo, A. (2016) 'Working at the web search engine side to generate privacy-preserving user profiles', *Elsevier on Expert Systems With Applications*, Vol. 64, pp.523–535.

Fang, H.Z. and Chen, C.C. (2015) 'A novel trend surveillance system using the information from web search engines', *Elsevier on Decision Support Systems*, Vol. 88, pp.85–97.

Hariharan, S., Dhanasekar, S. and Desikan, K. (2015) 'Reachability based web page ranking using wavelets', *Elsevier on Procedia Computer Science*, Vol. 50, pp.157–162.

Jeong, R.O., Oh, J., Kim, D.J., Lyu, H. and Kim, W. (2014) 'Determining the titles of web pages using anchor text and link analysis', *Elsevier on Expert Systems with Applications*, Vol. 41, No. 9, pp.4322–4329.

Khodambashi, S., Perry, A. and Nytrø, O. (2015) 'Comparing user experiences on the search-based and content-based recommendation ranking on stroke clinical guidelines – a case study', *Elsevier on Procedia Computer Science*, Vol. 63, pp.260–267.

Kim, N.H., Rawashdeh, M., Alghamdi, A. and Saddik, A.E. (2012) 'Folksonomy-based personalized search and ranking in social media services', *Elsevier on Information Systems*, Vol. 37, No. 1, pp.61–76.

Li, L., Xu, G., Zhang, Y. and Kitsuregawa, M. (2011) 'Random walk based rank aggregation to improving web search', *Elsevier on Knowledge-Based Systems*, Vol. 24, No. 7, pp.943–951.

Malthankar, S.V. and Kolte, S. (2016) 'Client side privacy protection using personalized web search', *Elsevier on Procedia Computer Science*, Vol. 79, pp.1029–1035.

Mishra, R., Kumar, P. and Bhasker, B. (2015) 'A web recommendation system considering sequential information', *Elsevier on Decision Support Systems*, Vol. 75, pp.1–10.

Nguyen, S.T.T., Lu, H.Y. and Lu, J. (2014) 'Web-page recommendation based on web usage and domain knowledge', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 10, pp.2574–2587.

Santos, O.C., Boticario, J.G. and Marín, D.P. (2014) 'Extending web-based educational systems with personalised support through user centred designed recommendations along the e-learning life cycle', *Elsevier on Science of Computer Programming*, Vol. 88, pp.92–109.

Shafiq, O., Alhajj, R. and Rokne, J.G. (2015) 'On personalizing web search using social network analysis', *Elsevier on Information Sciences*, Vol. 314, pp.55–76.

Wang, G.J., Huang, J.Z., Guo, J. and Lan, Y. (2015) 'Query ranking model for search engine query recommendation', *Springer on International Journal of Machine Learning and Cybernetics*, 2017, Vol. 8, No. 3, pp.1019–1038.

Wen, H., Fang, F. and Guan, L. (2012) 'A hybrid approach for personalized recommendation of news on the web', *Elsevier on Expert Systems with Applications*, Vol. 39, No. 5, pp.5806–5814.

Xu, Z., Liu, Y., Mei, L., Hu, C. and Chen, L. (2014) 'Generating temporal semantic context of concepts using web search engines', *Elsevier on Journal of Network and Computer Applications*, Vol. 43, pp.42–55.