Affinity Degree as Ranking Method

Rosyazwani Mohd Rosdan[1], Wan Suryani Wan Awang[2], Samhani Ismail[3]
Faculty Informatics and Computing, University Sultan Zainal Abidin, Besut Campus, Terengganu, Malaysia[1, 2]
Faculty Medicine, Medical Campus, University Sultan Zainal Abidin, Medical Campus, Terengganu, Malaysia[3]

***Abstract*—In machine learning, ranking is a fundamental problem that attempts to rank a list of things based on their relevance in a certain task. Ranking can be helpful, especially for future decision making. The framework for ranking has been classified into three primary approaches in machine learning: pointwise, pairwise, and listwise. However, learning to rank in all three approaches still lacks continuous learning ability, particularly when it comes to determining the degree of relevancy of ranking orders. In this paper, an affinity degree technique for ranking is proposed as another potential machine learning framework. The definition and attributes of the affinity degree technique are discussed, as well as the results of an experiment adopting the affinity degree approach as a ranking mechanism. The experiment's performance is measured using assessment metrics such as Mean Average Precision (MAP).**

Keywords*—*Affinity; affinity degree; rank; machine learning

I. Introduction

Learning to rank is a machine learning framework that aims to organise things in a particular order according to preference and relevance. Due to its emerging use in domains like information retrieval (IR) and recommender systems, learning to rank has drawn the attention of many machine learning researchers in the recent decade. The main reasons for the machine learning framework for ranking shared the exact nature of classification and regression methods. Also, the machine learning method can tune the parameters to overcome the disadvantages in the IR model, such as low precision and rigidness [1]. Learning to rank can be another predictive analytic technique under machine learning that presents learning to rank approaches [2]. Thus, learning to rank can be categorised as supervised learning with training and testing phases [3] and solving evaluation problems in search relevancy ranks [4]. Similar to other machine learning frameworks, the performance of learning to rank models is measured using the loss function that computes the difference between prediction and ground truth [5].

Dong, Chen, Guan, Li, and Xu mentioned the issues of learning to rank as a lack of continual learning ability and complicated tasks to construct a large-scale and resourceful training set [1]. Falah also mentioned the deficiency of current learning to rank approaches as lacking continual learning ability [4]. Therefore, this paper aims to incorporate the affinity degree classification algorithm into the rank technique as part of the learning ability for learning to rank issues. Since the ranking methodology also used classification and regression to rate the variables, the affinity degree classification algorithm might better fit the ranking system.

An affinity degree is a calculation for determining the degree of relationship and classification of the correlated data. Affinity degree has been established in peer-to-peer network data replication [6] as the calculation to define the similarity between two or more correlated data. The study used an affinity degree to find the correlation between files from different nodes. The correlation data is calculated to find the most binding factors contributing to the similarity between files. The results obtained from the calculation then will be ranked based on parameters. Therefore, affinity degree calculation has been implemented as one of the machine learning classification techniques in predictive analytic [7]. Thus, this paper will explore the affinity degree technique to rank as a machine learning framework.

The rest of the paper is organised as follows; Section 2 introduces the learning-to-rank theoretical background. After that, Section 3 describes more about affinity degree. Section 4 experiment for adopting the affinity degree into the learning-to-rank framework. Section 5 discussed the details of the experiment to validate the proposed idea. Finally, section 6 concludes the paper.

II. Theoretical Background

In traditional IR approaches, machine learning techniques were booming for the ranking problem, in which the learning-based method aimed to use labelled data for practical ranking function [8]. Learning to rank encompasses mainly supervised algorithms where the method uses machine learning techniques to train the model in a ranking task. Learning to rank was successfully applied to defect prediction to rank modules based on their defectiveness in software engineering. In test prioritization, this method can rank test targets based on a testing objective [9].

Learning to rank can be categorised into pointwise, pairwise, or listwise. For pointwise procedures, the approaches formed the model from the score assigned by users to individual objects. The yield rank is a collection of records with conventional scores. There is no reliance between training reports since the training reports are utilised independently [1], [10]. The simplest form, pointwise ranking, can be treated as classification or regression by learning the numerical rank views of documents as an absolute quantity [11].

The pairwise procedure learns by comparing two training objects and their given ranks or ground truth [12]. Trained by training samples as object pairs with independent variables and learning the classification (regression) model, two records are doled out in each pair with two relevance scores by individuals.

Nonetheless, only the match report dependence is considered, which implies that dependence between each report within the total rank cannot be considered entirely [1], [11]. The applicability of such methods is limited by the high computational cost of pairwise comparisons of user rated items in generating the training samples for the binary classifier [10].

The third procedure, listwise approaches, learn from the list of records. The records are relegated to a query in each list with diverse pertinence scores. Typically, this approach optimises a smooth approximation of a loss function that measures the distance between the references list of ranked items in the training data and the ranked list of items produced by the ranking model [10]. One common advantage is that more reliance between records is considered than pointwise and pairwise models with unreliable flexibility [1]. Meanwhile, Hass points out that the pairwise and the listwise approaches usually perform better than the pointwise approach [12].

Finding a suitable algorithm for a specific data set is significant for extracting the best information. Therefore, comparing the algorithm and ranking them into order will help indicate which algorithm should be applied. For selecting the best algorithm for a problem given, Carlos presents combination techniques called Zoomed ranking, which analyses the given data set and compares it with the relevant data set that has been processed by an algorithm using the "distance" concept for calculation [13]. Also, Bradzil presented three ranking methods: average rank, success rate ratio and significance win for algorithm selection [14]. The ranking methods eventually were being evaluated by average weighted correlation measures.

The ranking system has several different frameworks besides machine learning. Thus, there are various studies about the application of machine learning in ranking challenges and the importance and advantage of ranking in the machine learning framework. Yongyao Jiang addresses the ranking challenge in geospatial data discovery and proposes a system architecture to combine existing search-oriented open-source software, semantic knowledge base, ranking feature extraction, and machine learning algorithm [15]. Results show that the machine learning approach outperforms other methods in terms of both precision at K and normalised discounted cumulative gain.

Besides, the importance of machine learning rank or learning to rank in the construction of the IR system has been pointed out in [16]. Because each query has a set of associated documents represented by feature vectors that reflect the relevance of the documents to the query, it is a goal to build a model to predict the ground truth label of test data as accurately as possible in terms of the loss function. Also, it can be used to explore multiple ranking algorithms across different approaches in the item of accuracy and efficiency. Also, Hong Li specifically discussed exploring the fundamental problems existing approaches and future work in learning to rank [13]. Document retrieval is a task where the system maintains a collection of documents. The system retrieves the query words from the collection, ranks the document and returns the top-ranked documents.

Although ranking systems are most common in the IR environment, recent studies prove that the system can be applied in different environments, such as the medical field. The ranking system was used for ranking the Multimodal Features extracted from Congestive Heart Failure (CHF) and Normal Sinus Rhythm (NSR) subjects. Use high ranking features for detection of CHF and normal subjects. The findings indicate that the proposed approach with feature ranking can be beneficial for automatic detection of congestive heart failure patients and can be very helpful for clinicians and physicians' further decision-making to decrease the mortality rate [17]. A case study from Iran in which A Rad used the AHP algorithm and data mining to cluster and rank university majors [18]. Also, in the data mining field, D. Scully proposes an effective and efficient combined regression and ranking method that optimises the regression and objectives simultaneously [19]. Koshti used the learning to rank pairwise approach to making faster and better decisions for recruiting football players and having a list of options ranked on given criteria [2].

Nevertheless, regarding the affinity definition that is proposed to be used in this paper as a ranking technique, there is a study presenting a novel ranking scheme, Affinity Rank, which utilises two metrics [20]. The focus of the study is to evaluate the diversity of information retrieval performance. Measures the topic coverage of a group of documents, and information richness, which measures the amount of information contained in a document. Although the affinity in the rank system is not entirely new, there are not many of them.

III. Affinity Degree Method

Affinity is a notion that has received widespread attention in domains such as chemistry, biology, physics, social networks, security, and computer science. Affinity can hold a different meaning based on various concerns. Here, affinity is defined as a relationship, similarity, dependency and closeness between variables. Following is the affinity notation used in data replication by Awang [6]. The study proposed combining popularity and affinity files as the most critical parameters in replica selection. Affinity files were defined as the similarity between two or more correlated files before the system replicated the file. The affinity set is a set of any data that creates an affinity between files. Thus, the affinity between sets A and B consists of the intersection of elements between A and B plus the target and is not a null set. The equation can define the target in set B as fid (B), where f is a file and id refer to the file id.

$$aff_{AB} = \{x | x \in (A \times B + \{f_T(B)\}) \neq \emptyset\} \quad (1)$$

Definition 1: Let $A = \{f_{a1}, f_{a2}, \ldots . f_{an}\}$ and $B = \{f_{b1}, f_{b2}, \ldots . f_{bn}\}$, T is a targeted class. The sets A and B are said affinity denoted by (1); where f_T (B) is the target class in B.

$$aff_{AB}^{A} = \frac{aff_{AB}}{A + f_T(B)} \quad (2)$$

Definition 2: The affinity degree between A and B concerning A is defined as (2). The value expresses the degree of affinity between the data set A, and the affinity sets AB concerning A.

IV. Experiment

The main idea of affinity degree implementation is to measure the dependency or correlation between cause and particular effect. Measurement results might predict the set with the highest affinity degree as the leading cause of that effect. Therefore, this experiment focused on defining the risk of which symptoms can lead to a heart disease diagnosis. Through the affinity degree results, where the value of affinity degree was classified into five classes based on a specific indicator, the experiment could analyse the probability by ranking the affinity strength or assuming the correlation between dependent and independent variables.

This experiment was conducted according to KDD process in Fig. 1 [21]. Start with data selection, preprocessing and transformation data as the process results were shown in Table I, then affinity degree equation implementation. The ranking results were displayed in Tables II, III, IV and V by categories before the evaluation process.

In this experiment, MAP will be used as an evaluated method. MAP stands for the mean of the average precisions for each query computed. Average precision is computed as the sum of precisions for each found and relevant document, divided by the number of relevant documents. Using this construction, relevant but not found objects receive a precision of zero [22].

A. Heart Disease Data

The heart disease datasets used in this research were obtained from the Heart Disease Databases in the UCI Machine Learning Repository [23]. This data set dates from 1988 and consists of four databases contributed by the Cleveland Clinic Foundation (CCF), Hungarian Institute of Cardiology (HIC), Long Beach Medical Center (LBMC), and University Hospital in Switzerland (SUH), respectively. Each heart disease database has the same clinical instance format for each patient with 76 attributes, including the target attribute. It consists of 1025 patients with 499 patients ruled with heart disease while 526 were healthy. The target field refers to the presence of heart disease in the patient. It is an integer, valued at 0 or 1, indicating the absence or presence of coronary heart disease in patients. For other attributes, the integer, valued from 0 to 4, stated heart disease's absence, presence, and severity. Several risk factors can be controlled and cannot be controlled. The risk factors that can be controlled are blood pressure, blood cholesterol level, smoking, diabetes, obesity, inactivity and stress.

Meanwhile, a risk factor that could not be altered was age, gender, family history, and race. As part of preprocessing, this paper's attributes were compared to significant risk factors mentioned in the previous study for simplicity. Pre-processing focused on the handling of missing values, discretisation of numeric attributes and removal of instances with missing values [24]-[25]. Later, the attribute was compared with Hajar [26], Berg Gundersen, Sørlie and Bergvik [27], Mack and Gopal [28] and McClelland et al. [29]. For the age attribute, the age class was divided through class interval where the highest age minus the lowest age before was divided with the number of classes. Table I shows the reduced attribute details used in this experiment. Also, to get a better analysis, the data then were clustered into four categories: male older, male younger, female older, and female younger.

The experiment then implemented the adaptive equation in Section 3 defined as (2) into the data set. The affinity degree value then was ranked from highest to lowest displayed in Tables II, III, IV and V. The rank results show that the symptoms for each category were different. So, gender and age might greatly influence indicating the risk factor for heart disease diagnosis. For evaluating, this experiment used MAP as a tool, where the results will be discussed more in the next section.

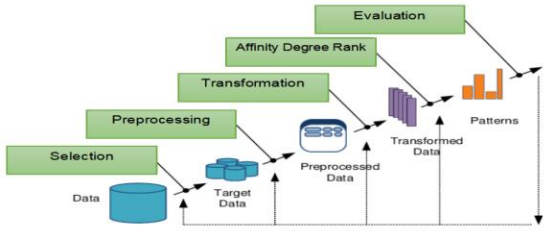

Fig. 1. Diagram of KDD Process.

TABLE I. Heart Disease Attribute List and Description

Attribute	Descriptions
age	0 = (<40)
	1 = (40-59)
	2 = (60-79)
	3 = (≥80)
gender	0 = female
	1 = male
Chest pain type (cp)	0 = typical angina
	1 = atypical angina
	2 = non-angina
	3 = asymptomatic
Resting blood pressure in mm Hg (trestbps)	0 = (<120)
	1 = (120-129)
	2 = (130-139)
	3 = (≥140)
Total cholesterol in mg/dL (chol)	0 = (<200)
	1 = (200-239)
	2 = (≥240)
Fasting blood sugar > 120 mg/dL (fbs)	0 = false
	1 = true
Maximum heart rate (thalach)	0 = (<60)
	1 = (60-100)
	2 = (>100)
Presence of heart disease	0 = absence
	1 = presence;

Table II displays the affinity degree rank for the female who is an older category. There are only 9 patients in this category with 0.811 for the highest and 0.801 for the lowest affinity degree values. Meanwhile, the younger female category shown in Table III has 0.858 for the highest and 0.802 for the lowest with 45 patients in this category. Table IV shows the affinity degree values for an older male category with 19 patients. The highest value for affinity degree in this category is 0.814, while the lowest is 0.802. Last, Table V with 90 patients for the younger male category shows the highest affinity degree values are 0.873, and the lowest is 0.802.

TABLE II. AFFINITY DEGREE RANK FOR OLDER FEMALE CATEGORY

age	gander	cp	trestbps	chol	fbs	thalach	Class	AffinityDegree
O	F	0	Hyper 2	Border	1	C	Abs	0.811638591
O	F	2	Hyper 2	High	0	C	Pre	0.808306709
O	F	2	Normal	High	0	C	Pre	0.808306709
O	F	1	Elevated	High	0	C	Pre	0.806709265
O	F	3	Hyper 2	Border	0	C	Pre	0.806709265
O	F	0	Normal	Desirable	0	C	Pre	0.803514377
O	F	0	Normal	Border	0	C	Pre	0.801916933
O	F	2	Elevated	Border	0	C	Pre	0.801916933
O	F	2	Hyper 2	Desirable	0	C	Pre	0.801916933

TABLE III. AFFINITY DEGREE RANK FOR YOUNGER FEMALE CATEGORY

age	gander	cp	trestbps	chol	fbs	thalach	Class	AffinityDegree
Y	F	0	Hyper 2	High	0	C	Abs	0.858
Y	F	0	Hyper 2	High	0	C	Pre	0.851
Y	F	0	Hyper 1	High	0	C	Abs	0.839
Y	F	0	Hyper 1	High	0	C	Pre	0.832
Y	F	0	Normal	High	0	C	Abs	0.825
Y	F	2	Elevated	Border	0	C	Pre	0.821
Y	F	1	Hyper 1	Border	0	C	Abs	0.821
Y	F	0	Normal	High	0	C	Pre	0.818
Y	F	0	Elevated	Border	0	C	Abs	0.818
Y	F	2	Hyper 2	High	0	C	Pre	0.816
Y	F	2	Normal	High	0	C	Pre	0.816
Y	F	0	Hyper 1	High	1	C	Abs	0.816
Y	F	1	Hyper 1	High	1	C	Abs	0.816
Y	F	1	Hyper 1	Border	0	C	Pre	0.813
Y	F	1	Hyper 1	High	0	C	Pre	0.813
Y	F	0	Hyper 2	Border	0	C	Abs	0.812
Y	F	0	Hyper 2	Border	1	C	Abs	0.812
Y	F	0	Hyper 3	High	1	C	Abs	0.812
Y	F	2	Hyper 1	High	0	B	Abs	0.812
Y	F	2	Hyper 1	High	0	C	Pre	0.812
Y	F	0	Elevated	Desirable	0	C	Abs	0.810
Y	F	0	Hyper 2	Desirable	0	C	Abs	0.810
Y	F	0	Elevated	Border	0	C	Pre	0.810
Y	F	0	Hyper 1	Desirable	0	C	Pre	0.810
Y	F	1	Hyper 1	High	1	C	Pre	0.808
Y	F	3	Hyper 2	High	1	C	Pre	0.808
Y	F	0	Elevated	High	0	C	Pre	0.807
Y	F	0	Hyper 1	Border	0	C	Pre	0.807
Y	F	1	Elevated	High	0	C	Pre	0.807
Y	F	1	Normal	Border	0	C	Pre	0.807
Y	F	1	Normal	Desirable	0	C	Pre	0.807
Y	F	2	Elevated	High	0	C	Pre	0.807
Y	F	2	Normal	Desirable	0	C	Pre	0.807
Y	F	2	Hyper 1	Border	0	C	Pre	0.804
Y	F	1	Hyper 2	Desirable	0	C	Pre	0.802
Y	F	1	Hyper 2	High	0	C	Pre	0.802
Y	F	2	Elevated	Desirable	1	B	Pre	0.802
Y	F	2	Hyper 1	Desirable	0	C	Pre	0.802
Y	F	2	Hyper 1	High	1	C	Pre	0.802
Y	F	2	Hyper 1	High	1	C	Pre	0.802
Y	F	2	Hyper 1	High	1	C	Pre	0.802
Y	F	2	Hyper 2	Border	0	C	Pre	0.802
Y	F	2	Hyper 2	Desirable	0	C	Pre	0.802
Y	F	2	Hyper 2	High	1	C	Pre	0.802
Y	F	2	Normal	Border	0	C	Pre	0.802

TABLE IV. AFFINITY DEGREE RANK FOR OLDER MALE CATEGORY

age	gander	cp	trestbps	chol	fbs	thalach	Class	AffinityDegree
O	M	0	Elevated	High	0	C	Abs	0.814701378
O	M	1	Hyper 2	High	0	C	Abs	0.814701378
O	M	2	Hyper 2	High	0	C	Abs	0.814701378
O	M	0	Elevated	Border	0	B	Abs	0.811638591
O	M	0	Elevated	Border	0	C	Abs	0.811638591
O	M	0	Hyper 1	High	0	C	Abs	0.811638591
O	M	0	Hyper 2	Desirable	0	C	Abs	0.811638591
O	M	0	Hyper 2	High	0	C	Abs	0.811638591
O	M	0	Normal	Border	0	C	Abs	0.811638591
O	M	2	Hyper 2	Border	0	C	Abs	0.811638591
O	M	0	Elevated	High	1	C	Abs	0.810107198
O	M	0	Hyper 2	Desirable	1	C	Abs	0.810107198
O	M	0	Normal	High	0	C	Abs	0.810107198
O	M	2	Hyper 2	High	1	C	Abs	0.810107198
O	M	0	Elevated	High	0	C	Pre	0.806709265
O	M	1	Hyper 2	High	0	C	Pre	0.806709265
O	M	0	Hyper 2	Border	0	C	Pre	0.803514377
O	M	2	Normal	High	0	C	Pre	0.801916933
O	M	3	Hyper 2	Border	1	C	Pre	0.801916933

TABLE V. AFFINITY DEGREE RANK FOR YOUNGER MALE CATEGORY

age	gander	cp	trestbps	chol	fbs	thalach	Class	AffinityDegree
Y	M	0	Hyper 2	High	0	C	Abs	0.873
Y	M	0	Hyper 2	High	0	C	Pre	0.867
Y	M	0	Elevated	High	0	C	Abs	0.862
Y	M	0	Hyper 2	Border	0	C	Abs	0.859
Y	M	0	Normal	Border	0	C	Abs	0.859
Y	M	0	Elevated	High	0	C	Pre	0.856
Y	M	0	Hyper 2	Border	0	C	Pre	0.853
Y	M	0	Normal	Border	0	C	Pre	0.853
Y	M	1	Elevated	High	0	C	Abs	0.848
Y	M	0	Normal	High	0	C	Abs	0.842
Y	M	1	Elevated	High	0	C	Pre	0.842
Y	M	0	Hyper 1	High	0	C	Abs	0.839
Y	M	0	Normal	High	0	C	Pre	0.835
Y	M	0	Elevated	Desirable	0	C	Abs	0.833
Y	M	0	Hyper 1	High	0	C	Pre	0.832
Y	M	2	Hyper 1	Border	0	C	Abs	0.830
Y	M	0	Hyper 1	High	1	C	Abs	0.828
Y	M	0	Elevated	Desirable	0	C	Pre	0.826
Y	M	1	Hyper 1	Border	0	C	Pre	0.826
Y	M	3	Hyper 2	High	0	C	Abs	0.825
Y	M	2	Hyper 2	Border	0	C	Abs	0.824
Y	M	2	Hyper 1	Border	0	C	Pre	0.823
Y	M	0	Elevated	High	0	B	Abs	0.822
Y	M	0	Hyper 2	Desirable	0	C	Abs	0.822
Y	M	2	Elevated	High	0	C	Abs	0.822
Y	M	2	Normal	Border	0	C	Abs	0.822
Y	M	0	Elevated	Border	0	C	Abs	0.821
Y	M	0	Hyper 1	Border	0	C	Abs	0.821
Y	M	0	Hyper 1	Desirable	0	C	Abs	0.821
Y	M	0	Normal	Desirable	0	C	Abs	0.821
Y	M	2	Normal	High	0	C	Abs	0.821
Y	M	2	Hyper 1	High	0	C	Pre	0.818
Y	M	3	Hyper 2	High	0	C	Pre	0.818
Y	M	2	Elevated	Border	0	C	Abs	0.818
Y	M	3	Hyper 1	Border	0	C	Abs	0.818
Y	M	2	Hyper 1	Desirable	0	C	Pre	0.816
Y	M	2	Hyper 2	Border	0	C	Pre	0.816
Y	M	0	Normal	Border	1	C	Abs	0.816
Y	M	1	Hyper 2	Border	0	C	Abs	0.816
Y	M	1	Normal	Border	0	C	Abs	0.816
Y	M	2	Elevated	Desirable	0	C	Abs	0.816

Y	M	2	Hyper 1	High	1	C	Abs	0.816
Y	M	2	Hyper 2	High	0	C	Abs	0.816
Y	M	2	Normal	Desirable	0	C	Abs	0.816
Y	M	3	Elevated	Border	0	C	Abs	0.816
Y	M	2	Hyper 2	Desirable	0	C	Abs	0.815
Y	M	0	Hyper 2	Desirable	0	C	Pre	0.815
Y	M	1	Elevated	Border	0	C	Pre	0.815
Y	M	2	Elevated	High	0	C	Pre	0.815
Y	M	2	Normal	Border	0	C	Pre	0.815
Y	M	0	Elevated	Border	0	C	Pre	0.813
Y	M	0	Hyper 1	Border	0	C	Pre	0.813
Y	M	2	Normal	High	0	C	Pre	0.813
Y	M	0	Elevated	Border	1	C	Abs	0.812
Y	M	0	Elevated	High	1	C	Abs	0.812
Y	M	0	Hyper 2	Border	1	C	Abs	0.812
Y	M	0	Hyper 2	High	1	C	Abs	0.812
Y	M	2	Elevated	Border	1	C	Abs	0.812
Y	M	3	Hyper 1	High	1	C	Abs	0.812
Y	M	1	Hyper 1	High	0	C	Pre	0.812
Y	M	2	Elevated	High	1	C	Pre	0.812
Y	M	0	Hyper 2	Desirable	1	B	Abs	0.810
Y	M	0	Hyper 2	High	0	B	Abs	0.810
Y	M	1	Hyper 3	High	0	C	Abs	0.810
Y	M	3	Normal	High	0	C	Abs	0.810
Y	M	1	Elevated	Desirable	0	C	Pre	0.810
Y	M	2	Elevated	Border	0	C	Pre	0.810
Y	M	0	Normal	Border	1	C	Pre	0.808
Y	M	1	Hyper 2	Border	0	C	Pre	0.808
Y	M	1	Normal	Border	0	C	Pre	0.808
Y	M	2	Elevated	Desirable	0	C	Pre	0.808
Y	M	2	Hyper 1	High	1	C	Pre	0.808
Y	M	2	Hyper 2	Border	1	C	Pre	0.808
Y	M	2	Hyper 2	High	0	C	Pre	0.808
Y	M	2	Normal	Desirable	0	C	Pre	0.808
Y	M	3	Elevated	Border	0	C	Pre	0.808
Y	M	3	Normal	Desirable	0	C	Pre	0.808
Y	M	1	Elevated	Border	1	C	Pre	0.807
Y	M	2	Hyper 2	Desirable	0	C	Pre	0.807
Y	M	2	Hyper 2	Desirable	1	C	Pre	0.807
Y	M	1	Normal	High	0	C	Pre	0.804
Y	M	3	Hyper 2	Border	0	C	Pre	0.804
Y	M	1	Normal	Desirable	1	C	Pre	0.802
Y	M	2	Hyper 1	Desirable	1	C	Pre	0.802
Y	M	2	Hyper 2	High	1	C	Pre	0.802
Y	M	3	Elevated	Desirable	0	C	Pre	0.802
Y	M	3	Hyper 2	Border	1	C	Pre	0.802
Y	M	3	Hyper 2	Desirable	0	C	Pre	0.802
Y	M	3	Hyper 2	High	1	C	Pre	0.802
Y	M	3	Normal	Border	0	C	Pre	0.802

V. Evaluation and Discussion

The experimental results reveal a variance of affinity degree that shows the relations or correlation between data with various affinity degree values. Shown in Fig. 2, the affinity degree rank differs in mean average precision between each category, and the differences are just a small gap. With 0.39 for the male and older category, the second category for female and older had 0.40, 0.27 for the third category, male and younger, and the last category for female and younger, with 0.29 in value of mean precision. For overall mean average precision, the value is 0.34.

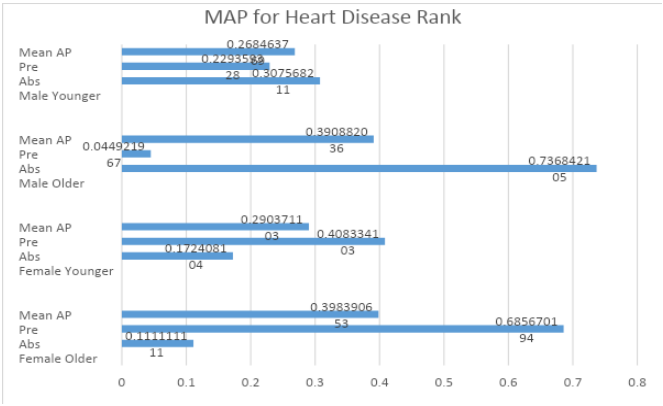

Fig. 2. The Evaluation Result of MAP for Heart Disease Rank.

All the value for mean average precision in each category were less than 0.5. The number of instances in each category might influence the evaluation results. For example, in male and older category, there are 14 instance of presence and only 5 for absence instances. Therefore, the gap between these two instances were small. Same goes to male and younger category, although the total instances were 90, but the gap between two instances were only 6. The small gap between instances does influence the mean average precision calculation.

The affinity degree is calculated to determine the relationship between heart disease symptoms and the diagnosis. From the coronary heart disease data sets, all 1025 records of patients were taken for calculation purposes. The data set was clustered into four groups according to the patient's gender and age. From the affinity degree calculated in this experiment, the highest score of degree or rank can be the most potential attribute for the patient to be diagnosed with heart disease or not. The limitation in this experiment were the results are not verified as there is no domain expert were involved. In future, more experiments with with variance data volumes need to be done along with the domain expert verified the results.

VI. Conclusion

This paper implemented the notion of affinity as another alternative technique for the ranking system. Heart disease experiments with an enhancement of the affinity degree equation have been done. The experiment defines the strength of correlation or dependency between data then ranks them based on affinity degree value. The experiment was evaluated by the MAP method, which uses the mean of average precision to compute for a set of queries. The results have shown the potential of affinity degree as one of the rank techniques. More experiments for diverse data samples with larger data volumes could be used to validate and verify the equation in the future.

Acknowledgment

Thanks to the internal grant of UNISZA (UniSZA/2021/DPU2.0/08) for financially supporting our work. Also, thanks to all team members for reviewing for spelling errors and synchronisation consistencies and for the constructive comments and suggestions.

References

[1] X. Dong, X. Chen, Y. Guan, S. Li and Z. Xu, "An overview of learning to rank for information retrieval," in 2009 WRI World Congress on Computer Science and Information Engineering, IEEE, March 2009, vol. 3, pp. 600-606, doi: 10.1109/CSIE.2009.1090.

[2] A. V. Koshti, "Learning to Rank Model Performance and Review with Pairwise Transformations," M.S. thesis, Creative Components, 757, Iowa State Univ., Ames, Iowa, 2021.

[3] H. Li, "A short introduction to learning to rank," IEICE TRANSACTIONS on Information and Systems, vol. 94(10), pp. 1854-1862, Oct. 2011, doi: 10.1587/transinf.E94.D.1.

[4] F. Al-akashi, "Learning-to-Rank: A New Web Ranking Algorithm using Artificial Neural Network," International Journal of Hybrid Innovation Technologies, vol.1(1), pp.15-32, 2021, doi: http://dx.doi.org/10.21742/ijhit.2021.1.1.02.

[5] A. Rahangdale and S. Raut, "Machine learning methods for ranking," International Journal of Software Engineering and Knowledge Engineering, vol. 29(06), pp. 729-761, 2019, doi: 10.1142/S0218194019300001X.

[6] W. S. W. Awang, M. M. Deris, O. F. Rana, M. Zarina and A. N. M. Rose, "Affinity replica selection in distributed systems," in International Conference on Parallel Computing Technologies. Springer, Cham, Aug. 2019, pp. 385-399, doi: https://doi.org/10.1007/978-3-030-25636-4_30.

[7] R. M. Rosdan, W. S. W. Awang and W. A. W. A. Bakar, "Comparison of affinity degree classification with four different classifiers in several data sets," International Journal of Advanced Technology and Engineering Exploration, vol. 8(75), pp. 247, 2021, doi: http://dx.doi.org/10.19101/IJATEE.2020.762106.

[8] M. F. Tsai, T. Y. Liu, T. Qin, H. H. Chen and W. Y. Ma, "Frank: A ranking method with fidelity loss," in Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, Jul. 2007, pp. 383-390.

[9] A. Bertolino, A. Guerriero, B. Miranda, R. Pietrantuono and S. Russo, "Learning-to-rank vs ranking-to-learn: strategies for regression testing in continuous integration," in Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering, June 2020, pp. 1-12, doi: https://doi.org/10.1145/3377811.3380369.

[10] J. Liang, J. Hu, S. Dong and V. Honavar, "Top-N-Rank: A Scalable List-wise Ranking Method for Recommender Systems," in 2018 IEEE International Conference on Big Data (Big Data), IEEE, Dec. 2018, pp. 1052-1058.

[11] H. Valizadegan, R. Jin, R. Zhang, and J. Mao, "Learning to Rank by Optimising NDCG Measure," NIPS Vol. 22, pp. 1883-1891, Jan 2009.

[12] R. Haas and B. Hummel, "Learning to rank extract method refactoring suggestions for long methods," in International Conference on Software Quality, Springer, Cham, Jan. 2017, pp. 45-56, doi: 10.1007/978-3-319-49421-0_4.

[13] C. Soares and P. B. Brazdil, "Zoomed ranking: Selection of classification algorithms based on relevant performance information," in European conference on principles of data mining and knowledge discovery, Springer, Berlin, Heidelberg, Sept. 2000, pp. 126-135.

[14] P. B. Brazdil and C. Soares, "A comparison of ranking methods for classification algorithm selection," in European conference on machine learning, Springer, Berlin, Heidelberg, May 2000, pp. 63-75.

[15] Y. Jiang, Y. Li, C. Yang, F. Hu, E. M. Armstrong, T. Huang,... and C. J. Finch, "Towards intelligent geospatial data discovery: a machine learning framework for search ranking," International journal of digital earth, vol. 11(9), pp. 956-971, 2017, doi: http://dx.doi.org/10.1080/17538947.2017.1371255.

[16] Learning to rank. Retrieved from https://www.cs.purdue.edu/homes/liu1740/report.pdf.

[17] L. Hussain, W. Aziz, I. R. Khan, M. H. Alkinani and J. S. Alowibdi, "Machine learning based congestive heart failure detection using feature importance ranking of multimodal features," in Mathematical Biosciences and Engineering, vol. 18(1), pp. 69-91, 2020, doi: 10.3934/mbe.2021004.

[18] A. Rad, B. Naderi and M. Soltani, "Clustering and ranking university majors using data mining and AHP algorithms: A case study in Iran," in Expert Systems with Applications, vol. 38(1), pp. 755-763, 2011, doi: 10.1016/j.eswa.2010.07.029.

[19] D. Sculley, "Combined regression and ranking," in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Jul. 2010, pp. 979-988.

[20] Y. Liu, B. Zhang, Z. Chen, M. R. Lyu and W. Y. Ma, "Affinity rank: a new scheme for efficient web search," in Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, May 2004, pp. 338-339.

[21] Monitoring Online Tests through Data Visualization - Scientific Figure on ResearchGate. Retrieved from: https://www.researchgate.net/figure/The-Steps-of-a-KDD-process_fig7_220073492 [accessed 24 Mar, 2022].

[22] Trotman, A. (2005). Learning to rank. Information Retrieval, 8(3), 359-381.

[23] UCI Machine Learning Repository. Heart disease data set. Retrieved from https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease.

[24] R. Rosly, M. Makhtar, M. K. Awang, M. I. Awang, and M. N. A. Rahman, "Analysing performance of classifiers for medical

datasets," International Journal of Engineering & Technology, vol. 7(2.15), pp. 136-138, 2018, doi: 10.14419/ijet.v7i2.15.11370.

[25] M. Makhtar, R. Rosly, M. K. Awang, M. Mohamad and A. H. Zakaria, "A Multi-Classifier Method based Deep Learning Approach for Breast Cancer," Int. J. Eng. Trends Technol, (1), pp. 102-107, 2020, doi: 10.14445/22315381/CATI3P217.

[26] R. Hajar, "Risk factors for coronary artery disease: historical perspectives," Heart views: the official journal of the Gulf Heart Association, vol. 18(3), pp. 109, 2017.

[27] A. E. Berg Gundersen, T. Sørlie and S. Bergvik, "Women with coronary heart disease–making sense of their symptoms and their experiences from interacting with their general practitioners," Health Psychology and Behavioral Medicine, vol. 5(1), pp. 29-40, 2017.

[28] M. Mack, A. Gopal, "Epidemiology, traditional and novel risk factors in coronary artery disease," Heart failure clinics, vol. 12(1), pp. 1-10, 2016.

[29] R. L. McClelland, N. W. Jorgensen, M. Budoffff, M. J. Blaha, W. S. Post, R. A. Kronmal and A. R. Folsom, "10-year coronary heart disease risk prediction using coronary artery calcium and traditional risk factors: derivation in the MESA (Multi-Ethnic Study of Atherosclerosis) with validation in the HNR (Heinz Nixdorf Recall) study and the DHS (Dallas Heart Study)," Journal of the American College of Cardiology, vol. 66(15), pp. 1643-1653, 2015, doi: http://dx.doi.org/10.1016/j.jacc.2015.08.035.