the republic of Iraq

ministry of higher education and scientific research

university of Al-Qadisiyah

college of computer science and information technology

computer department

Web User profiling in java using semantic matching

A report submitted to the department of computer science of the requirements for obtaining a bachelor's degree in computer science and information technology / computer department

2017 2018

Set by:

Nijod Faisal Abd Al-Sattar

Haneen Thamir Kadhim

Supervisor:

Dr. Muntasir Jaber Jawad

بســـم الله الرحمن الرحيـــم

"اقرأ بأسم ربك الذي خلق ✺ خلق الأنسان من علق ✺ اقرأ
وربك الأكرم ✺ الذي علم بالقلم ✺ علم الأنسان ما لم يعلم

صدق الله العلي العظيـــم

"اهداء"

جل وعلا شأنه امرنا بالقراءة واكدها مرتين . الاولى انارت الطريق والثانية لأناره الطريق للأخرين . انا اقرأ وهم يحموني ويعلموني الحشد الشعبي المقدس واستاذي الدكتور منتصر جابر جواد .

Chapter one

Introduction

1-XML

XML is a shortcut to (extensible markup language). It designed for transmitting data and stored. Some believe that XML used to display the data as in HTML but not, and is the markup language general to create language coding with your purpose have the ability to describe a number of different types of data that means it's way to describe the data as in the data base. XML is used as a format to store and process documents that are connected and not connected to the internet.

Currently there are two version of XML:

The first version is XML (1.0), which appeared in 1998 and is now in it's fourth version, which appeared in 2006 and it widely used.

The second version is XML (1.1), which appeared in 2004 and is now in it's second version, which appeared in 2006 and it is not widely used:

Characteristics of the XML which makes the language of appropriate for data transfer.

1-The formula humanity and read automatically.

2-Her support for characters international standard system that allows for any information in any language written connect.

3-Its ability to represent the most common computer science data structure such as list and tree.

4-It has a self-documentation format that describes the structure, domain names, and assigned values. [1]

2-Matching

Matching: is the process of discovering mapping between two graphs through the application of matching algorithm, there are two types of matching: [7]

2.1-Exact Matching

the exact matching method is for improving the estimation of causal effects by reducing imbalance in covariates between treated and control groups. The exact method is faster and easier to used and understand, requires fewer assumption, more easily automated, and processes more attractive statistical properties for many applications than existing matching methods. In exact matching, user temporarily coarsen their data, exact match on these coarsened data then run their analysis on the coarsened, it bounds the degree of model dependence and causa effect

estimation error by extant user choice, it is reducing the maximum imbalance on one variables has not effect on others, doesn't require a separate procedure to restrict data to common support meets the congruence principle, is approximately invariant to measurement error and it is enable to balances all nonlinearities and it is interaction in simple and can works with multiply imputed data set. In the other type of matching methods inhered many of CEM's properties when applied to further match data preprocessed by exact matching. [2]

2.2-Semantic Matching

Semantic matching is a new approach and discussed some of it's key properties. For performing generic matching. We search for semantic correspond by mapping meaning (concepts), and not labels, as in syntactic matching. When we match two nodes, it is not sufficient to consider the meanings of labels of this node, but also need the positions of the nodes in the graph. The semantic is the similarity relations between elements (concepts) rather than the syntactic similarity. Compute element-level semantic matching for each node, compute semantic relation holding among all concepts denoted by labels at nodes under consideration. [7][8]

3-User Profiling

The user profile is represented by using XML language. We represent it using XML because it has many features in representation profile information. Such as data and documents represented in XML can be processed with different types of applications, that makes it more applicable switch and sharing data or documents between components (application, database).

The user profile contains both general user information that is applicable over all applications and more unstable application particular data the user profile data is approved against an XML pattern to secure the integrity of the data. User profile is showing personal data associated with specific user or customized desktop environment. A profile indicates therefore to the explicit digital representation of a person's identity. A user profile also can be considered as the computer representation of user form. A profile can be used to store description the characteristics of a person. This information can be used to by systems taking into account the persons characteristics and preferences. [6]

Chapter two

Literature review

Related works

1-According to the work of "Anna Formica":

The method that aimed at finding the best matches between a user request and service offered by several enterprises, that suggest a given business ecosystem (for example, the tourism sector) a group of SMEs (SMEs it is represent the backbone of the European industrial system) agree on the adoption of reference ontology. It is used to build the company profiles depending on the services provided. Thus; a user gives set of desired features the represent user required, is expressed in terms of the reference ontology terminology (concepts). Explanation of SemSim, a method used to collectively search the SME profiles to identify the services that match at better the user required. SemSim is the approach that used to evaluated the semantic similarity among concepts depending on the well-known information content. The similarity assessment is analyzed by studying the correlation among the selected similarity methods and human judgment. The correlation reflects the noisiness in the linear relationship among a human judgment and for instance semsim values that essentially means that higher

scores on HJ tend to be paired with higher scores on semsim, analogously for lower scores. [3]

2- According to the work of "Changbo ke":

We depend a web service modeling ontology discovering frame work. Through the proposing conception similarity and structure similarity based on taxonomic and a hierarchical methodology, we avoid a complex logic reasoning effectively and by defining a serial of restructuring constraints according to the relationship among two similarities and restructuring. With development of web service standard and a maturity of platform that support the web service development that leads to make the web service a major software paradigm and computing the resource. To describe web service by web service description language that based on XML. It includes grammatical compatibility but lacks the consideration of semantic information. In addition, the service registration and discovery mechanism is based on the global description of its discovery and generation and supports grammatical operation This level has two double points of this service to describe the service cannot describe the service more accurately, on the other hand, the services are obtained only by the presence of a corresponding word in the process of

discovery of the service and therefore difficult to satisfy the user requirements in the functional and non-functional Internet services And be unable to detect the services of the agent or user accurately but the features of ontology web service, semantic web service it's smart to accurately describe the service and increase the efficiency of discovery

The results of this work are to increase the accuracy of the description of the services of users to the Internet sites and to increase the efficiency of discovery of these services.

[4]

Finally, in work of Anna Formica, we have been able to get the results of intensive and good for medium enterprises and get the results confirmed, but we will not be able to use this work in large enterprises or large businesses and control of small and medium companies are a small part of our work in comparison to our other needs in the field of control.

And in work of Changbo Ke, we were able to cut down the services provided by users and increase the efficiency of their discovery of application submission by users but we could not contain all the requirements of users in this work, but we controlled a small part of the most important of them.

Chapter three

Suggested Method and Algorithms

1-Similarity Distance

We depend in this topic to the language of java.

User profile between the two persons by using the language of java depending on the XML for any two people, for example, first person "A" and second person "B" using the same as the search engine. We compared them and find out the behavior oh each them. At follow-up to the behavior of "A" we see that the most word used in the search engine (for example, computer) and he used (1000 times) and "B" also use the word (computer) and they are the most commonly used to him in the same search engine and used (100 times) in this case when the comparison among them can we say they are close behavior and distance similar among them (whenever less than the distance increased similarity).

2- Contextual matching

To enhance the semantic expression of the texts along with the traditional keyword matching strategy so as to effectively improve the contextual matching in our approach we take two aspects of the similarities among pages and ads.

1- **Similarity based on keywords and common text capture**

2- Similarity based on Wikipedia measure the relevance of the semantic perspective of oneness-like techniques. This approach consists of the following steps:

- First: -of all we choose enough articles from Wikipedia to share many of the semantic concepts.
- Second: -we build the keyword phrase for each page. Finally, we suggest combining the two types above and similar in a uniform way. To make the top-nods selection.
- In order to evaluate the effectiveness of our approach we conducted a set of experiments containing real ads and pages, the results show that the approach combines Wikipedia based semantic matching with word matching can greatly improve the accuracy of the measurement of similarity among pages and ads and thus improve the effectiveness of contextual advertising. In addition, the results of full-text matches take along time between al articles and pages.[5]

Term Frequency Inverse Document Frequency (TF-IDF) Algorithm:

We study TF-IDF to identify words in a set of documents that are more suitable for use in the query. TF-IDF refers to the calculation of the value of each word found in the documents through the inverse proportion of the frequency

of words in the specified parts of the documents and calculating the percentage of them. -IDF means the strength of the relationship with the specific documents indicating that if these words appear in the query, this document is useful to the user as a result of his request. [9]

The example of TF-IDF Algorithm: -

```
package com. guendouz. textclustering. preprocessing;
 import java. util.Arrays;
 import java. util. List;
 public class TFIDFCalculator {
public double tf(List<String> doc, String term) {
double result = 0; for (String word: doc) {
if (term. equalsIgnoreCase(word)) result++;} return result /
doc. size ();
}
public double idf(List<List<String>> docs, String term) {
double n = 0;
for (List<String> doc: docs) {
for (String word: doc) {
 if (term. equalsIgnoreCase(word)) {
 n++;
 break;
 } } }
return Math.log (docs. size () / n);
 }
 public double tfIdf(List<String> doc, List<List<String>>
docs, String term) {
 return tf (doc, term) * idf (docs, term);
}
public static void main(String[] args) {
List<String> doc1 = Arrays.asList("Lorem", "ipsum",
"dolor", "ipsum", "sit", "ipsum");
List<String> doc2 = Arrays.asList("Vituperata",
"incorrupte", "at", "ipsum", "pro", "quo");
List<String> doc3 = Arrays.asList("Has", "persius",
"disputationi", "id", "simul");
 List<List<String>> documents = Arrays.asList(doc1, doc2,
doc3);
 TFIDFCalculator calculator = new TFIDFCalculator ();
double tfidf = calculator. tfIdf (doc1, documents, "ipsum");
System.out.println("TF-IDF (ipsum) = " + tfidf);
 }
 }
```

Chapter four
Plans for next steps

In the beginning of November, I started to write the introduction and it took a period of no more than two weeks. I dealt with the topics that show the XML and the methods of dealing with them and Matching and the types that include Semantic matching and Exact matching and how they work and benefit from them and then dealt with the subject user profile and its uses in various aspects and how to learn from him.

Then, at the end of November, I began to write a literature review which lasted almost a month and illustrated the work that was similar to the work I did and the results of these works by the people and the characteristics and weaknesses of each of these acts and give her personal opinion.

Then I wrote the suggested methods and algorithms which were made up of three main parts. The first part was to explain the similarity distance and the second part to contextual matching. The third part was an explanation of the TF-ID algorithm and its mode of operation.

After this work I plan to do programs that monitor the user interfaces and know the sites visited by the remote and take advantage of them know user sites favorite and send ads and proposals suitable for them and future work to secure this work and make it safer than the present.

Chapter five

Conclusion

1- **XML (extensible markup language)**

It designed for transmission data and stored. There are two version of XML

- **The first version is XML (1.0)**
- **The second version is XML (1.1)**

2- **Matching: -is the process of discovering mapping among the graphs through the application of matching algorithm, and there are two types of matching**

- **Exact matching**
- **Semantic matching**

3- **User profiling: -is represented by using XML language. User profile is showing personal data associated with specific user or customized desktop environment.**

4- **TF-IDF Algorithm: -calculates values for each word in a document through an inverse rate of the frequency f the word in a particular document to the percentage of documents in which they appear. The words that contain high (TF-IDF) numbers. Means a strong relationship with documents showing which indicates that if the word appearing to the user. This algorithm effectively classification related words that can enhance query retrieval.**

References

1- **XML طريقك الى**

2- **User profiling in intrusion detection**

3- **Semantic search for matching user requests with profiled enterprises**

4- **Self-adaptive semantic web service matching method**

5- **Improving contextual advertising matching by using Wikipedia thesaurus knowledge**

6- **User profile management reference model and web services implementation**

7- **Semantic matching**

8- **Java platform, Enterprise edition (Java EE) specification, v7**

9- **Using TF-IDF to determine word relevance in document queries**

Contents

Contents