Deema and Jawad, J Inform Tech Softw Eng 2018, 8:2
DOI: 10.4172/2175-7866.1000227

Journal of
Information Technology & Software Engineering

Research Article Open Access

Similarity Calculation Framework for Heterogeneous Multimedia Web Resources

Deema A and Jawad B*

College of Computer and Information Sciences, King Saud University Riyadh, Saudi Arabia

Abstract

With the huge amount and diversity of multimedia resources available on the web, it is becoming necessary to satisfy user's information and knowledge needs in the mobile computing era. These resources need to be described and correlated systematically to allow applications and web services to access and reuse them properly. Although popular web search engines are very efficient in retrieving any type of information they are not designed to search different types of multimedia information at the same time. Traditional information retrieval techniques should be enhanced by semantic techniques, which analyze web resources information and metadata and use similarity calculations to find correlations between any two resources on the web. Web resources metadata are the attribute values that characterize a multimedia file and give a description about its content. In this research we propose a method that measures the semantic similarity between multimedia resources in the web by using resources' metadata attributes.

CCS concepts: Information Systems---Information Retrieval---Retrieval Models and Ranking ---Similarity Measures.

Keywords: Information retrieval; Similarity calculation; Multimedia metadata; Web resources similarity

Introduction

Today, the spread of social media applications in the world in addition to the new mobile technology allowed people to communicate more easily. Social media become part of our daily activities. Right now, everyone can upload any information on the web as well as retrieving it. The process requires only few seconds. This makes the available information on the web increasing radically. The web become like a big repository of resources and information, and because of this there is a need to describe efficiently these resources for retrieval purpose later.

Although web search engines are very fast and efficient in retrieving information, these are not designed to deal with different types of resources at the same time. This limitation forces users to use the same search engine many times to search for different types of multimedia resources such as videos, audios, presentations, maps, apps, etc. Indeed, when the user triggers a search, it is done for a specific type of resource, other types are ignored other resources might be more related and better satisfy the user query than the returned results. Combining different types of multimedia resources in the same search may be very useful for users who are looking for the best multimedia match for their queries and would be a valuable support for many usages such as e-learning authoring and course design, social media information retrieval, knowledge management, multimedia information search.

This paper proposes the use of similarity techniques to improve traditional information retrieval. These similarity techniques use similarity calculations to find correlations between any two resources on the web by using metadata, which describes the resource on the web. Metadata is a collection of attributes that describe specific multimedia object or resource. Each type of multimedia object has a certain set of attributes. Images, videos, audios and documents each have a fixed set of descriptive metadata. For example, Image can be described by a title, camera model, creator, dimensions and so on. Metadata provides a semantic meaning of the content which makes calculating the similarity between any types of multimedia files straightforward. Experiments done on a set of heterogeneous multimedia resources show that cosine similarity outperforms other similarity measures in matching metadata associated with the resources

Background and Related Work

Information retrieval overview

There have been many developed models to retrieve information. These models calculate a numeric score on how well each resource in the database matches the query and then rank the objects depending on this score. The top ranking objects are then shown to the user. The process may then be repeated if the user wants to revise the query. The following are the most used models for similarity calculation in IR.

Boolean model: 1) Some users find it difficult to structure effective Boolean queries. 2) AND, OR and NOT, have a different meaning when used in a query, So when users use the natural language terms to form a Boolean query, they tend to make errors because they revert to their knowledge of English. Also, the binary nature of the retrieval decision in Boolean systems is frequently cited as a drawback because the Boolean Model is not able to rank the returned list of documents [1,2].

Statistical model: The second IR model is the Statistical Model, which includes the vector space and the probabilistic retrieval approach. Both of these two models use statistical information and term frequency to find the documents that are relevant to the query. They both use the term frequency in a different way, but they both have the same output, which is a list of documents ranked by their value of relevance. This model has some advantages over the traditional Boolean model. It allows computing a continuous degree of similarity between queries and documents, and so it can rank the documents based on their probable relevance, and also allows partial matching [3].

***Vector space model*:** The Vector Space model considers the

***Corresponding author:** Jawad B, College of Computer and Information Sciences, King Saud University Riyadh, Saudi Arabia, Tel: +91-80-28612445/46; E-mail: d3.s@live.com

Received February 04, 2018; **Accepted** March 04, 2018; **Published** March 10, 2018

Citation: Deema A, Jawad B (2018) Similarity Calculation Framework for Heterogeneous Multimedia Web Resources. J Inform Tech Softw Eng 8: 227. doi: 10.4172/2175-7866.1000227

documents and the query as vectors embedded in multidimensional Euclidean space, the dimensions are used to represent the index of documents. This model ranks the documents based on the similarity between the query and the documents. In the VSM, the query and the document represent as vectors, each dimension represents a particular term, if the term included in the document then its value score in the vector space equals non-zero. There are several ways to calculate these values or (term weight) one of them is tf-idf weighting [4].

VSM used the term frequency as a weighting factor, it was proposed by Salton, Wong and Yang [5]. This model is known as: (Term frequency–inverse document frequency), in short tf-idf weighting. The (tf) factor measures the frequency of term occurrence in the document or query, and the (idf) factor measures the inverse of the number of documents that contain a query term.

***Probabilistic model*:** Robertson and Jones [6] developed the Probabilistic relevance model. Given a query, this model attempts to ranks all the returned documents based on the probability of their relevance to the given query. Also, it assumes that there is a set, from the retrieved documents, that is preferred by the user, this set called R which means "document is relevant" and all the remains documents that not presented in R are denoted as ▪ "document is not Relevant" [3].

$$sim(d_j, q) = \frac{P(R|\vec{dJ})}{P(R|\vec{dJ})} \qquad (1)$$

Linguistic and knowledge-based model: The third model is the Linguistic and Knowledge-based model. In the simple form of information retrieval, the user enters some keywords string, which is used to search the inverted indexes of the keywords included in the document. The simple traditional model retrieves documents based only on the existence of exact single word strings that given in the query. So, obviously, this simple approach fails to capture some relevant documents because it does not understand the complete or deep meaning of the user's query [2]. The Linguistic and Knowledge-based model have been adopted to solve this problem by using a morphological and some semantic analysis to provide different meaning, to retrieve documents more effectively.

Semantic similarity overview

Semantic similarity definition: Semantic Similarity measures the semantic equivalence between two linguistic terms rather than measuring the similarity of their syntactical representation (e.g. string format). These two items either be concepts, documents, or sentences. Semantic Textual Similarity (STS) is measuring the similarity between documents or sentences [7,8].

Many applications in Natural Language Processing field (NLP) Artificial Intelligence (AI), and cognitive science and psychology consider Semantic Similarity as an essential component to compute similarity between two objects. For human, the most popular way to compare between two things is to find the similarity between these two items. Comparing two words together is an easy job for people, for example: given three words: Car, Apple and Automobile, it's easy to recognize that car and automobile are correlated somehow but Apple doesn't have any connection to them [9].

Similarity measures: Similarity measures are necessary to understand many pattern recognition issues such as in information retrieval, clustering and classification. There are many distance/similarity measures that can be used to compare two objects, either semantically or syntactically. From a mathematical side, distance is a quantitative degree of how objects are far from each other. Distance means dissimilarity, so we can measure the similarity between two objects by measuring how close these objects are to each other [10]. Here we are focusing on the most famous measures to find the similarity:

***Cosine similarity*:** Cosine similarity is VSM based method that measures the textual similarity. To measure the similarity between two texts, the two texts must be converted to two vectors and then measure the cosine of the angle between them. Each term is assigned a particular dimension, and the document is represented as vector where the value of each dimension depending on the number of time that term occurs in the document. The similarity cosine function is defined as:

$$sim(d_j, q) = \frac{dj.q}{\| dj \| \| q \|} = \frac{\sum_{i=1}^{N} W_{i,j} W_{i,q}}{\sqrt{\sum_{i=1}^{N} W_{i,j}^2} \sqrt{\sum_{i=1}^{N} W_{i,q}^2}} \qquad (2)$$

Documents and queries are represented as vectors.

$$q = (w_{1,q}, w_{2,q}, ..., w_{n,q})$$

$$q = (w_{1,q}, w_{2,q}, ..., w_{n,q})$$

The resulting similarity ranges from 1 which means (exactly the same), to zero, which means not related or (dissimilar).

***Euclidian distance L2*:** Euclid stated that the line is the shortest distance between any two points, this is known as Euclidean distance. Since it was derived from Pythagorean theory, it is often referred to as the 'Pythagorean distance' [11].

If x and y are two vectors, start from zero coordinates, in the Euclidian space, the distance d between x and y is calculated by:

$$d(x, y) = d(y, x) \sqrt{(x1 - y1)^2 + (x2 - y2)^2} \qquad (3)$$

[x1, x2] is the coordinates of vector x which origin is zero (0,0)

[y1, y2] is the coordinates of vector x which origin is zero (0,0)

And from this equation, we can calculate the distance in any n-dimensional space:

$$X = [X_1, X_2 ...., X_n]$$

$$Y = [Y_1, Y_2 ...., Y_n]$$

$$d(x, y) = \sqrt{(x1 - y1)^2 + (x2 - y2)^2 + ... + (xn - yn)^2} \qquad (4)$$

***Manhattan distance L1*:** Manhattan distance was named after the grid layout of Manhattan's streets. The distance between two points is the sum of the difference of their Cartesian coordinates.

$$d(x, y) = \sum_{i=1}^{n} |xi - yi| \qquad (5)$$

Where (x, y) are vectors:

$$X = [X_1, X_2 ...., X_n] \text{ and } Y = [Y_1, Y_2 ...., Y_n]$$

***Jaccard coefficient*:** Jaccard Similarity or Jaccard index is a method to compare between two sets of data. The Jaccard similarity calculated

by divided the number of common features by the size of the union of the data set. [12].

Jaccard sim (A, B) =(P (A∩B))/(P (A∩B)) (6)

Related works

This section presents some previous studies related to implementing metadata and semantic similarity in information retrieval. There are not so many studies in this domain, but we tried to find some useful previous works because this will help us to understand the problem properly.

Spoerri A [13] presents the Cite4me application that helps the students in the learning process. This application implements a standard IR technique with the semantic web approach to search and recommend scientific publications for users. It used a LAK (Learning Analytics and Knowledge) dataset, which is a collection of metadata of the scientific papers that was published in some known conferences. The dataset contains 315 descriptions of papers, including information about the authors, the conference, and the content of the papers. The similarity between matching papers and non-matching papers in the LAK dataset was calculated by using cosine similarity measure, which is used after computing the tf-idf scores. Furthermore, recommendation feature in this system was implemented by calculating the distance between given entities to define the relatedness between documents, then ranking the retrieved documents based on the final score.

Cha SH [14] represents an attempt to use metadata in document management system. The authors built a prototype system called an AWOCADO (Adaptive Workflow Controller and Document Organizer). It provides a novel framework for managing and retrieving documents by using their attributes or metadata. The information gathered from the document or the metadata was stored in a repository called Metadata Store. The repository held about 700 documents and almost 5000 metadata. The experiment was conducted for three months. The goal of this system was to lectures hosted by Western Kentucky University. This repository contains thousands of different types of online lectures such as: PowerPoint, video, text, podcast, and audio. This platform works as the main online connection between the university and the students. So the authors implement a metadata specific search engine and compare the results with the generic one. The implementation was done in four phases: first was extracting the knowledge of the resources and this step was done manually to extract the information for each lecture. The second step was adding the metadata information such as: the collage name, course name, lecture name, teacher, and format. Then they used Nutch open-source search engine. Nutch has different types of fields beside text such as, keywords or metadata, so the authors used this feature to add their own metadata fields that mentioned previously. The Nutch implementation is based on the cosine similarity of VSM and Boolean model.

Multimedia Resources Description

Metadata overview

Metadata is mostly known as "data about data" or a data that give information about other data. It is used to give a summarization or basic information about data, thus it can make tracking and retrieving the specific data or resource a lot easier. Some example of metadata include: date and time, author or creator, file size, keywords…etc.

For example, a video file could contain some metadata that describe the content of the video, the length of the file, the creator, the type of camera that was used to capture this video, the date and time when the video was created and so on. Also, there is a metadata for the web pages on the Internet, which can describe the content of the page, as well as key words linked to the content. These links are known as Meta tags, which used to be the main factor to arrange the retrieved results for a web search before 1990s [15].

Since the enormous increase of digital content, metadata have become an interested and important concept. Because at the end, it's more efficient and time saving to looks up the content or resources based on the information or metadata that has been linked to the resource rather than searching the content itself.

From the similarity point of view, if two objects or resources have a similar metadata, then they are probably very similar or related. For example: if two images have the same title they are probably similar. Thus, it's important to consider the metadata as separate information and not like the usual text content.

Metadata types for multimedia

Metadata was initially used in the card catalog for libraries to help those who are looking for a specific book. But as information start to transform to digital form, metadata also has been used to describe the digital objects to facilitate the retrieval of relevant information.

There are two types of multimedia metadata: automatic metadata and manual metadata. The first type is automatically collected metadata from software or device like the video camera. It stores the camera-related information such as lenses, aperture, shutter speed…etc. While the manual metadata is the information that users provide to describe the multimedia content, which can be a description, set of keywords, or some comment.

In this research the work will be focused on video files, images, documents, audio and presentations (slides).

Here are some of the metadata that can help to compute the semantic similarity between these web resources:

Image metadata: Most of the digital image on the web doesn't contain only the picture, but also include some metadata about the image. Different image format will contain different type of metadata, but here are some common and basic metadata that can be exist with any digital image. Table 1 shows the main image metadata.

Video metadata: Most of the metadata for video created automatically. But the manually written metadata, by users, proves to be very important in recommending related content. YouTube today

Attribute	Image	Video	Document	Audio	Presentation
Title	✓	✓	✓	✓	✓
Category	✓	✓	✓		✓
Description	✓	✓			✓
Keywords	✓	✓			✓
Create	✓	✓	✓	✓	✓
date					
Location	✓	✓			
Device model	✓				
Author	✓	✓	✓	✓	✓
Subject			✓		
Genre				✓	
Album				✓	

Table 1: Metadata attributes of file.

relies on metadata such as description and keywords to optimize video searching and indexing. Table 1 shows the YouTube metadata.

Document metadata: There are so many types of text document files such as Microsoft Word or PDF. These documents can contain some metadata. Some of the metadata is automatically generated, and others require the user who created the document to write it manually.

Audio metadata: Every music or audio file has some kind of metadata, where specific information about the file or the artist is stored. The Audio Engineering Society (AES) sets some standards to describe the digital audio files. Table 1 shows the main Audio metadata.

Presentation (slides) metadata: One of the most known resources for education on the web is the presentations or slides. Table 1 shows a presentation metadata.

Similarity Calculation for Multimedia Resources

The first step to this system is collecting some type of web resources. Then extract all the metadata from these files and store it in the database, which will contains different types of multimedia files with their attributes. The similarity is measured between resources by using the previous measures we mentioned earlier, which are, cosine similarity, Euclidian distance, Manhattan distance. By using all of these measures we can identify which one would give us the best result. To calculate the similarity between words in the text, we first need to use some of the natural language processing techniques such as, stemming and stop-words removal. These techniques transform all the words in the text form to simple words to make calculating the similarity more efficient.

For the experiment, a series of tests have been done for each file type. The precision and recall scores for the four similarity measures are calculated and averaged. The following is an example test case for a video submitted to the system. The video selected from YouTube (https://www.youtube.com/watch?v=- 2FfdcIn9Nw) is a tutorial about pilgrimage. Table 2 shows the video metadata extracted from YouTube.

According to a reviewer used in the experiment, there are 12 resources of different types that are relevant to the selected video: 6 images, 3 presentations, and 3 videos. The weighted cosine similarity gives the following results ranked from the most similar to the less similar. When analyzed by the reviewer, she stated that results no 1, 2, 6, 7, 8, 9, 10 were really relevant and they show either a video or a presentation about pilgrimage (Hajj) or some pictures that was taken in Mena or Arafat (two main steps) during pilgrimage. Other similarity measures used did not perform better in terms of number and relevance of resources matched.

System Implementation

The system has been implemented as a desktop application in Java language, and SQL Database. The first step was to extract the metadata from different web resources by using some API like YouTube and Flicker. Then we populated the database with 500 records of different file (100 records for each file type). Before start the calculation process we used some stemming word technique and stop word technique to remove any punctuation marks and preposition words.

Title	Hajj movie – a journey of love
Description	Hajj movie that explain the Hajj and how to do it
Location	8360 Rd No 9 Southern Ln, Al Mashair, Mecca
Created date	2013/11/27
Keywords	Hajj movie; Hajj; Mecca; Manasik; Islam;
Category	Religions
Language	English

Table 2: Video metadata.

1. Hajj - the Journey of Hearts (presentation) Similarity: 16.0%
2. The Hajj (presentation) Similarity: 13.0%
3. hajj (Pilgrimage to Makkah) (presentation) Similarity: 12.0%
4. Hajj 1436 (image) Similarity: 10.0%
5. Hajj - Mena (image) Similarity: 7.0000000000000001%
6. Makkah - Mena (image) Similarity: 6.0%
7. Hajj- Mena (image) Similarity: 6.0%
8. Hajj 1430 (image) Similarity: 6.0%
9. Learn How to Perform Hajj Step By Step: Complete Video (video) Similarity: 5.0%
10. What is The Muslim Holy Pilgrimage aka Hajj? (video) Similarity: 5.0%

Figure 1: Weighted cosine similarity results.

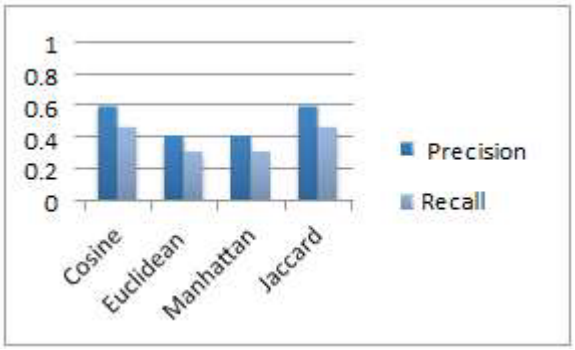

Figure 2: Precision and recall results.

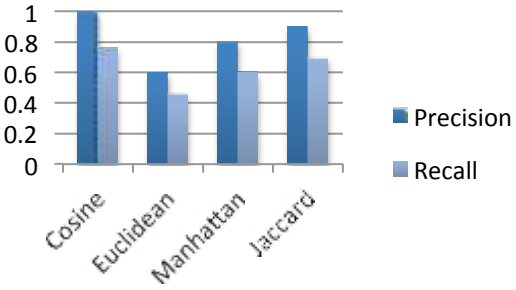

Figure 3: Weighting attributes results.

We tested the system with different type of files using different measures and the results were different with each measure function. We compute the quantitative analysis by using the precision and recall evaluation techniques. Also, to improve the effectiveness of the system performance we used some weighting factor for the most important attributes and the results really improved. Figures 1-3 represent the result of the experiments.

Conclusion

This research proposes a method to measure the similarity between heterogeneous multimedia resources in the web by using resources' metadata. This feature is becoming critical in search engines to allow retrieving any web resource that matches exactly the user's query. It

is also deeply needed for mobile users as multimedia information is the ideal way of communicating information through mobile devices. We tested four methods for measuring the similarity between web resources between different resource types. The results show that the cosine similarity gives the best results especially if it is combined with weighting factors. It is planned to extend the system by including many other types of web resources Also it will be interesting to test the system in real life as a tool that retrieves multimedia resources from existing web repositories.

References

1. Maind A, Deorankar A, Chatur P (2012) Measurement of semantic similarity between words: A survey. IJCSEIT 2: 189-194.
2. Fetahu B, Nunes BP, Casanova MA (2013) Cite4Me: Semantic retrieval and analysis of scientific publications. LAK-Data Challenge.
3. Hiemstra D, Vries PD (2000) Relating the new language models of information retrieval to the traditional retrieval models.
4. Jurgens D, Pilehvar MT, Navigli R (2014) SemEval-2014 Task 3: Cross-level semantic similarity SemEval. Proceedings of the 8th International Workshop on Semantic Evaluation, Dublin, Ireland.
5. Agirre E, Diab M, Cer D, Agirre AG (2012) A pilot on semantic textual similarity. Proceedings of the Sixth International Workshop on Semantic Evaluation Association for Computational Linguistics pp: 385-393.
6. William FB (1992) Information retrieval data structures & algorithms. Prentice Hall Inc, Upper Saddle River, NJ, USA.
7. Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Communications of the ACM, ACM New York, USA.pp: 613-620.
8. Salton G, McGill MJ (1983) Introduction to modern Information Retrieval. McGraw-Hill, New York.
9. Zhuhadar L, Nasraoui O, Wyatt R (2008) Metadata Domain- knowledge Driven Search Engine in "HyperManyMedia" E- learning Resources. CSTST: Proceedings of the 5th International Conference on Soft Computing as Transdisciplinary Science and Technology, New York, USA.
10. Andric MA, Hall W (2005) Using Metadata forI nformation Retrieval in Document Management Systems. Serbia & Montenegro, Belgrade.
11. Margaret R (2014) Metadata. TechTarget.
12. Zhang S, Zheng X, Hu C (2015) A survey of semantic similarity and its application to social network analysis. IEEE International Conference on Big Data. Santa Clara, CA, USA.
13. Spoerri A (1995) InfoCrystal a visual tool for information retrieval.
14. Cha SH (2007) Comprehensive survey on distance/similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Sciences 1: 301-307.
15. https://en.wikipedia.org/wiki/Jaccard_index