Collection System and Recommendation of Academic Texts with OCR Functions

Francisco Hidalgo Bueno
Instituto Politécnico Nacional, UPIITA, Av. IPN No.2580 Col. La Laguna Ticomán, Gustavo A. Madero. México, D.F., C.P. 07340México*
e-mail: hidalgo.bf@hotmail.com

Daniel Alfonso González Cervantes
Instituto Politécnico Nacional, UPIITA, Av. IPN No.2580 Col. La Laguna Ticomán, Gustavo A. Madero. México, D.F., C.P. 07340México*
e-mail: dan.el.gonzalez@gmail.com

Paola Nayeli Cortez Herrera*
e-mail: pcortez@ipn.mx

Carlos de la Cruz Sosa*
e-mail: carlosdelac@gmail.com

Itzamá López Yáñez*
e-mail: ilopezyb05@ipn.mx

Abstract—Today the search for information is a complex task, requiring an investment of time to find one that interests us. But having defined areas of interest is possible to focus the search more in order to optimize time. The aim of this study is to recommend thesis-based areas of interest for the user. For this, building the user profile is required, and in order not to limit the recommendation is considered the recommendations of other users. Algorithm it is implemented based on content and other in the collaborative model for recommendations. As additional part, the application through OCR will allow you find the thesis.

Keywords*-search, recommend thesis, user profile, OCR*

*****

I. Introduction

Currently, the amount of information available is considerable, accessing to a Web search engine will be enough to find millions of results. In addition to that, factors like information search intervene as the purpose of the search, which will have to discriminate the content to find what is really needed. Undoubtedly this process would require a large investment of time.

Recommendation systems make this process faster, although this does not relieve them of implementing certain strategies to efficiently handle the volume of documents that are delegated to not present the same problems as the physical storage repositories. Problems arise as a result of inadequate management of documents as the volume of these increases, causing difficulties for users to access the right information.
Imagine now as scenario a student looking for a thesis related to specific area of interest in a library. Even with the thesis classified by subject, it will be necessary to analyze in more detail which is the one that belongs to the area of interest that the user needs.In a few words, this process is not a simple task. Therefore this work focuses on developing a mobile application that retrieves stored thesis in a database, and provides recommendations based on the user's profile. The recommender system will be based on an IEEE taxonomy to specify the thesis search. It is considered that the application will have an optical module (OCR) which allows looking for information related to the advisers of the document by recognizing or obtaining information from the cover of the thesis becausecapturing information sometimes is a slow and tedious process.

Here it is described how the content of an article is structured. The introductory section provides an overview of the application and presents some basic concepts for the realization of the same, the framework includes the consulted work. The methodology is presented in section three and the diverse modules that make up the application, in the results section images of the application are shown. Finally conclusions are given.

A. Recommendation Systems

In a place where tasks are performed storage and management of documents for consultation purposes, such as libraries; It may become complicated the process of sorting the documents adequately.

Digital libraries and digital repositories may have similar problems, even if the searches are performed with tools provided by the site, the result yielded causes the end user losing significant information.

Recommender systems on the other hand, apply filtering techniques of information in order to facilitate access to users and guide them through the entire collection handled by the system elements that can meet those who really are linked to their interests.

Is possible to find these popular systems for various purposes in electronic commerce and scientific information services being. [5]

B. Optical Character Recognition

The main task of pattern recognition is to make the right decision to classify a pattern according to their characteristics (Fig 1). A complete procedure OCR (Optical Character Recognition) also adheres part of the pre-processing.

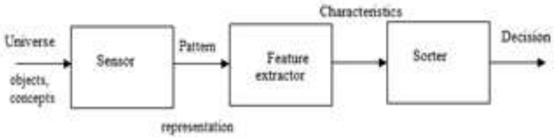


Figure 1. Basic functions of a text recognition system and forms

To extract useful information from the millions of pixels that has an image, the character recognition system must reduce the information that is provided to fewer selecting regions of interest and applying pre-processing tasks such as cleaning, improved dotted, etc. Thus we have a set of truly useful information for the purpose of recognition.

II. Framework

Some related works are presented with the proposal. In [3] it is shown the design of a model of document recommendation through the application of Semantic Web technologies for information management, use of RSS feeds is presented for generating custom newsletters, bibliographic alerts for each user (profiles) and modules to manage the creation of recommendations. The authors mention that when using an RSS feed is a greater visibility of the library on the web and wider dissemination of their resources. On the other hand in [7] it focuses on building profiles based on an analysis of natural language embodied in user reviews. The use of natural language for creating complex profiles can acquire information that goes beyond the simple weight or selecting an item. As a result of applying this methodology it is available a resource consisting of a vocabulary of words, phrases and expressions used to predict the record to which a given text belongs in mind. It is appreciated that both works have as main focus the recommendation of documents and classification of information, but through various approaches. This work tries to involve both approaches considering the first factor as the area of interest of the user to make the recommendation and as auxiliary factors the behavior of the user before the system.

In the commercial area you can appreciate companies that use the recommendation to assist its users, mentioned below:

Amazon. - American international e-commerce company, which provides recommendations for items based on search of history and other parameters of the user’s profile.[1]

Genius. - music recommendation service that is part of the iTunes online store, this application sends information about the songs in the user's personal library, comparing profiles of other members to recommend the collections of those with whom they have enough similarities. [4]

Daily Newspaper ME. - allowing virtual news recommendation from the personal tastes of the user. [2]

Last.fm. - social network that works around music recommendations, in addition it monitors musical events and the creation of forums. Their recommendations are based on comparing user data with the rest of the community through information-based radio stations to which the user accessed.[8]

Netflix.- is the leading Internet TV service with over 33 million members in 40 countries with approximately one billion hours of TV shows and movies per month including original series. The system provides a hint system of programs and allows the user to assign a grade that reflects their view of the show. [10]

LinkedIn.-The largest online professional network, was founded in May 2003 and has over 259 million members. This service allows you to contact professional in order to find work, other people and business opportunities. [9]

III. Methodology

For the development of the application was considered to split into sections each of the parties within it. In this section are explained in detail.

A. Taxonomy

The representation of the areas of knowledge used to classify both the theses of the system as user interests that will serve to generate the recommendations will be obtained from the IEEE 2013 v1.0 Taxonomy, whose structure is shown in Figure 2. [6]

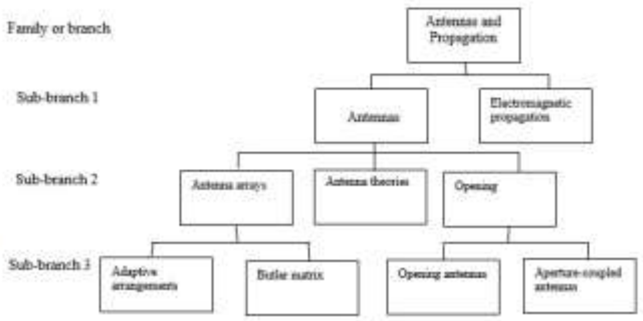


Figure 2. Partial representation of taxonomy. Antennas and Propagation. IEEE extract taxonomy.

B. User Profile

The user profile in the system basically consists of two groups of information (Fig 3). The first one, personal information, allowing to sketch other usersan idea who the person behind the profile is and his account. It consists of data such as name, age, gender, educational level and school of origin among a few others. As you might expect, they are provided by the same user who will decide which of them are going to be recorded and which are not.

The second group consists of information according to the user’s intererests which are collected by the application in two different ways. First, the same user provides a list of up to 6 areas of knowledge that are of interest, this is the fastest way the application uses to collect this type of information. Furthermore, the user has the possibility to make a number of theses and rate them as favorites from a range of one to five points (where one is the lowest score and five the highest). From this point it is possible to extract the areas reflected in each of the thesis of his activity (favorite and punctuated) knowledge.

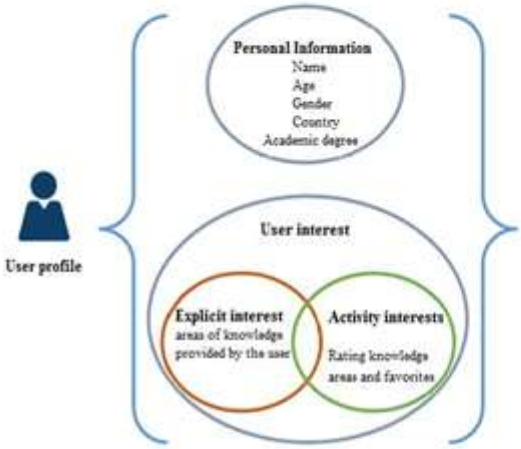


Figure 3. Composition user profile

For the composition of a profile, a matrix is used in which explicit user interests (determined by the area of interest and importance to the user assigned)are involved.

C. *Recommendation Algorithms*

A recommendation-based model on content is used to provide the active user (UA) specialized recommendations which elements to suggest keep a relationship of high similarity to the user's interest.

For making this posible, a list of areas of interest belonging to the AU are used, generated from the areas indicated by itself with those areas obtained from user's activity (thesis that have been the scored and added to favorites) it is made. This involves all areas related to the latter two branches arising from the same origin in the taxonomy of the IEEE (Fig. 2).

Additionally, the collaborative model is used because the application recommends the UA thesis related users or Registered users (UR), commented, rated and added to favorites.

For this, the level of similarity is calculated between the profile of the UA explicit interest profiles and other UR explicit interest, using Cosine UR Salton and those in which the results reflect an average of similarity (value selected between 0.45 and 0.55. It is probable that among those users elements to recommend are found, sothe AU will seek to develop new areas with the ones already recommended.Offering Multidisciplinary recommendations to expand the knowledge of the UA landscape.

D. *OCR text extraction*

OCR is built with the help of an API (Application Programming Interface) or function library that runs on the client side or Smartphone. Once a partial area of the cover of the thesis has been captured with the device camera, the image processing is performed to extract the text content in it (plain text) and send this information to the server that will use it to search for theses in the system and display information to the user.

In this model the full weight of the prosecution is in the mobile device, it is possible that there may be differences of time from the moment the user makes the catch until the application displays text on the screen depending on the processing power of the device.

In this case Tesseract, which is one of the most accurate OCR available free code in combination with leptonic Image Processing Library was used.

IV. Results

Below some images of the application with the results obtained are presented. First, there is the screen that forms the user's profile, where registration has explicit user interests. In Figure 4 it is shown how sections where the user enters information into the system to form his profile.are being classified.

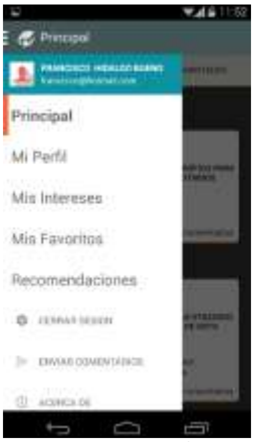


Figure. 4. Screen showing the user's profile in the application.

The taxonomy is triggered when the user accesses to "My Interests" (Fig. 4) section, and there, the branches of the taxonomy are shown, its objective is to find the subject of interest of the user. Also the user can enter the topic of interest and the application will display related information.

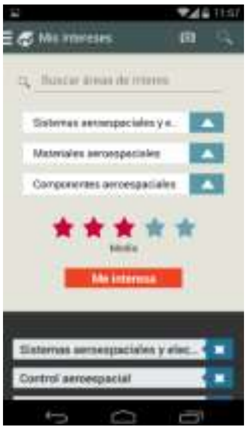
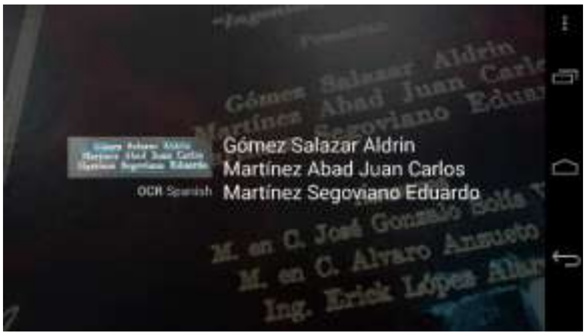

11:57
Mis intereses
Buscar áreas de interés
Sistemas aeroespaciales y e...
Materiales aeroespaciales
Componentes aeroespaciales
★★★★★
Media
Me interesa
Sistemas aeroespaciales y elec...
Control aeroespacial

Figure 5. The taxonomy in the application. The display shows how the user will indicate the application area of interest according to the taxonomy.

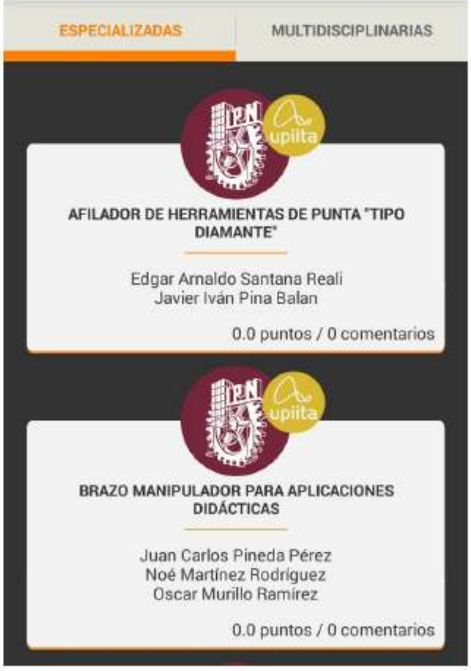
Other features already mentioned include the OCR module, when through the mobile phone camera a picture of the cover of the thesis is taken and the user can search the thesis.



Gómez Salazar Aldrin
Martínez Abad Juan Carlos
Martínez Segoviano Eduardo

Figure 6. Getting the names of the authors and / or consultants in a thesis, through OCR.

Briefly they have been presented through Figures 4 to 6 the functionalities of each of the modules of application. However the ultimate goal is as follows, where it is noted that once the user has a profile and gets involved in the application by comenting or scoring a thesis, is able to get recommendations.



AFILADOR DE HERRAMIENTAS DE PUNTA "TIPO DIAMANTE"

Edgar Arnaldo Santana Reali
Javier Iván Pina Balan

0.0 puntos / 0 comentarios

BRAZO MANIPULADOR PARA APLICACIONES DIDÁCTICAS

Juan Carlos Pineda Pérez
Noé Martínez Rodríguez
Oscar Murillo Ramírez

0.0 puntos / 0 comentarios

Figure 7. User has specialized or multidisciplinary input, according to what he is looking for.

V. Conclusions

Nowadays, due to the amount of information when seeking information, support tools are needed to aid the user in his search. The proposal presented in this paper is a mobile application that recommends bachelor's degree thesis level, the recommendation is based on a model for content and collaborative filtering. IEEE taxonomy was used in order to have a better classification of areas of interest, in addition to making use of optical character recognition (OCR) to help the searching or capturing information in an agile way. For testing a total of 125 users with their areas of interest were created, until now the algorithms have been successful. However it is needed to try in detail the collaborative algorithm, which the trial period will continue.

Acknowledgment

Acknowledgments. The authors thank the National Polytechnic Institute, UPIITA for the facilities provided for the realization of this thesis project.

.

References

[1] Amazon (2017) . 20 Enero 2017. https://www.amazon.com.mx
[2] Daily Newspaper (2017). 27 Enero 2017. http://www.nydailynews.com/services/faqs
[3] E. Peis, E. H. Viedma y J. M. M. d Castillo, <<Modelo de servicio semántico de difusión selectiva de información (DSI) para bibliotecas digitales>>, 2008. Thesis
[4] Genius (2017). 10 Enero 2017. https://genius.com/
[5] HiperText.net (2014). Universitat Pompeu Fabra. 11 Marzo 2014, http://www.upf.edu/hipertext/numero-6/recomendacion.html
[6] Institute of Electrical and Electronics EngineersIEEE (2017). 15 Abril 2014.<<IEEE Advancign Technology for Humanity,>>http://www.ieee.org/documents/taxonomy_v101.pdf.
[7] J. A. Roberto, M. A. Marti, P. Rosso, <<Sistemas de recomendación basados en Lneguaje Natural: opiniones vs valoraciones,>> 2011. Thesis
[8] Last fm (2017). Services. 15 Enero 2017, de Last fm. Sitio web: http://www.last.fm/api
[9] LinkedIn. (2016). Developers. 28 Enero 2017, de LinkedIn Sitio web: https://developer.linkedin.com/
[10] Netflix. (2016). Help. 28 Enero 2017, de Netflix Sitio web: https://help.netflix.com/en/node/412