Platform to promote research in Ecuador using Linked Data and Data mining

Nombre del Primer Autor[1], *Nombre del Segundo Autor*[2], *Nombre del Tercer Autor*[1,3]

[1]Afiliación del primer autor, nombre de la universidad,
Dirección de la Universidad, ciudad, pais, código postal

[2]Afiliación del segundo autor, nombre de la universidad,
Dirección de la Universidad, ciudad, pais, código postal

[3]Afiliación del tercer autor, nombre de la universidad,
Dirección de la Universidad, ciudad, pais, código postal

Corresponding author: {primero,segundo}@universidad.edu, tercero@universidad2.edu

Fecha de recepción: 21 de Septiembre, 2014
Fecha de aceptación: 17 de Octubre, 2014

ABSTRACT

(i) mencionar los principales objetivos y el alcance de la investigacion (lo que se hizo, por que se lo hizo y para quien se escribio el articulo?); (ii) describir los metodos empleados en la investigacion; (iii) resumir los resultados obtenidos; y (iv) mencionar las principales conclusiones derivadas de la investigacion. Los resumenes se redactan, por lo general, en tiempo pasado, porque se refiere al trabajo ya efectuado.

Keywords: *English Keywords, LATEX Template, Revista Maskana, University of Cuenca, DIUC.*

1. INTRODUCTION

La investigacion en Iberoamerica ha aumentado en los ultimos anos. Segun la publicacion del estado de la Ciencia del ano 2015 la cantidad de articulos registrados en Science CItation Index (SCI) [1] crecio en 123%. Aumentando su participacion en bases de datos internacionales al incrementar su produccion cientifica local [2]. Uno de los paises mas destacados es Brasil que aumento su cantidad de publicaciones en un 2.5. Sin embargo, se tiene varias limitaciones, como la cantidad de recursos invertidos en investigacion en contraste con el promedio mundial. Se debe tener en cuenta que Latinoamerica tiene el segundo crecimiento mas rapido del mundo despues de Asia [3]. Tiene una gran diversidad de areas de conocimiento, ademas, cada pais tiene diferentes formas estrategicas de abarcar los problemas de una region. Lo cual brinda un conjunto de soluciones que puede resultar una ventaja en comparacion a los paises de primer mundo, en el campo de la investigacion, ya que dichas soluciones deben ser capaces de lidiar con la heterogeneidad presente en Latinoamerica.

Uno de los enfoques de las IES (Instituciones de Educacion Superior) de Latinoamerica es contribuir en el desarrollo sostenible de la sociedad, mediante la cooperacion de sus alumnos y docentes, impulsados por la investigacion. Actualmente, cierta informacion acerca de investigadores y sus recursos bibliograficos se encuentran esparcidas entre varios repositorios digitales o bases de datos bibliograficas. Cuando se necesita proponer proyectos con varios investigadores de una area especifica que pertenecen a diferentes IES, surgen preguntas como: Quien trabaja en lineas de investigacion parecidas? o Como se puede crear una red de investigadores de un area en comun, cuando no conocemos si estos existen? Ademas, para definir el perfil investigativo de una persona en analisis, obtener sus articulos, conocer revistas en las que fueron aceptados, entre otros, es necesario acceder a varias fuentes de datos. Tomando en cuenta que este proceso es manual, sintactico y diferente por cada fuente de recursos bibliograficos disponible en la Web.

Ampliar el alcance de esta base de conocimiento a toda Latinoamerica permitira a los sistemas de educacion superior de nuestra region contar con un repositorio digital centralizado con informacion sobre recursos bibliograficos de investigadores ecuatorianos. Con este proyecto se pretende incentivar la colaboracion interinstitucional y asi obtener como fruto de este trabajo un repositorio semantico validado, con una herramienta para la localizacion de investigadores de areas similares de investigacion que proveera informacion actualizada. Potenciando asi la generacion de redes de investigacion con pares academicos en la region y brindando a las instituciones participantes mayores oportunidades de cooperacion y colaboracion .

En este documento se presenta una plataforma desarrollada que permite detectar areas de conocimiento similares entre investigadores y ayudar a formar grupos de trabajo interinstitucionales. Con el uso de enfoques orientados a la integracion de bases de datos bibliograficas disponibles en Internet como: Google Scholar, Microsoft Academics , Computer Science Bibliography (DBLP) , Scopus . Utilizando tecnologias de Web Semantica y procesos de descubrimiento del conocimiento (Knowledge Discovery in Databases o KDD).

The rest of this paper is organized in the following way:

2. RELATED WORK

Es necesario contar con herramientas que faciliten el trabajo a los investigadores, en el que varios proyectos han trabajado, como: Semantic Scholar[4] y [5]. Sin embargo estas herramientas abarcan un dominio limitado, es decir, son herramientas que tratan publicaciones y autores locales o de una determinada area de conocimiento como por ejemplo informatica. La estabilidad de las herramientas que traten este problema es fundamental ya que cada dia surgen nuevos aportes cientificos, nuevas areas de estudio, y cientos de investigadores se suman en proyectos tanto locales, como internacionales y multidisciplinarios.

3. THEORY

3.1. Text clustering

Document or Text Clustering [6] is a subgroup of the data clustering field [5] which is an unsupervised learning process. Clustering consists in organize items from a collection into groups of similar items, where each group is called cluster. Each cluster have a set of similar items to each other - generally based in a measure of similarity - but dissimilar to other items that belongs to other clusters. Clustering should not be confused with classification, because documents does not have a class assigned. Documents in text clustering are represented as a bag of words, which give a problem of high dimensional spaces. [1] There is no way to know the number of clusters, size or shape before to apply clustering. A human or an algorithm are who determine these parameters. [2,3].

Clustering is not as simple as it seems, to do more than just grouping, we could produce a disjoint (exclusive clustering) or overlapping partitions. Clustering algorithms could be divided in two flavours, discriminative and generative types. Discriminative algorithms are based on a distance metric to find a similarity between documents. While, generative algorithms the model is squeezed to fit in the distribution to produce cluster centroids. Finally clustering has been used in many areas like information retrieval [7], outlier detection [4], to improve queries returned by search engines [7], etc.

3.2. K-Means

3.3. Topic Model

3.4. Latent Dirichlet Allocation

4. Consolidate Data

Scientific publications of Ecuadorian authors are available in different bibliographics sources on the Internet at each source varies the information on scientific activity of an author. For example Scopus bibliographic sources as recorded affiliation of authors, tables, graphs of publications, authors study areas, etc. DBLP features that source does not cover. For this reason it is necessary to make a unification of these bibliographic resources of different disciplines, structures with features that feed a common data model. For this task has been defined as the first phase the extraction of scientific publications from different external literature sources, once you have the data integration process and disambiguation of authors and their publications is done, and as a third phase a method is defined data update add information to allow a controlled way, facilitating access, discovery and reuse of library resources. The development of the first stage of the platform faces two challenges. 1) Extracting data from heterogeneous bibliographic sources. 2) Integration of publications with different data formats, vocabularies and conceptualizations using ontologies and vocabularies for describing bibliographic data in a single model. Removing publications from bibliographic databases. The extraction process of publications is responsible for obtaining information from scientific articles Ecuadorian authors from various external bibliographic sources previously analyzed as Google Scholar, Microsoft Academics, Scopus, etc. Each of these data sources operate in a different way, so the extraction service adapts to each of the data sources. For the extraction of publications should be considered that the data source consisting of an API access, because if data collection is not dare an API the data quality is poor as demonstrated below. The collection of scientific publications is the first phase that will allow us to obtain data which depend on the following phases, because if the data are erroneous alter the expected results.

5. Detect Similar Areas

In this section, we outline the web service built for data processing . The service has been called KODAR that means ”Discovery Of Knowledge Research Areas”with the words a little bit jumbled. It uses Apache Mahout to execute algorithms of machine learning. We choose mahout for the ability to deal with massive datasets, it is a scalable Java library and we could profit of the distributed computation, because It is built upon Apache Hadoop. KODAR has three main stages that are: Discover similar areas, detect researchers networks and find a general topic area. The whole implementation is open sourced and available on our GitHub repository.

5.1. Discover similar areas

Broadly, keywords of academic literature talk about a certain topic area or methodology. Detecting similar areas based in the keywords . It could help us to detect researchers with interests in common and open up an opportunity to generate new research projects. Boosting interagency collaborative work and form cooperative research groups.

Firstly, we disjoin our data, because we just need to process the keywords to detect similar areas. Other fields like author or title of the publication are stored in a separate file. Both files are converted in a specific Hadoop file format that is SequenceFile[1]. This file stores key/value pairs, where the key is a unique identifier and a bunch of keyword that belong to a paper are stored as a value. Same happens in the another file with the difference in the value pair. We store the remaining fields.

It is necessary to do some procedures before to clustering the data into Mahout. Data has been preprocessed to convert text in numerical values, but not all the keywords have the same relevance. The weighting technique used to magnify the most important words is Term frequency-inverse document frequency (TF-IDF). The weighted values are used to generate the Vector Space Model (VSM) where words are dimensions. The problem with this VSM generated is that words are entirely independent each other and It is not always true. Sometimes words have some kind of dependency like Semantic with Web. In order to achieve this dependency we use collocations. At the time of writing, we are executing our experiment using bi-grams and an Euclidean norm (2-norm), which can change. In future experiments, It will be interesting to generate vectors using Latent Semantic Indexing (LSI) or apply a log-likelihood to take words that mostly have the chance to go together. So in the long run, we have our vectors completed to clustering.

We start with the vectors generated to execute K-Means algorithm in Mahout. It was executed using a Cosine distance measure as the similarity measure. RandomSeedGenerator[2] was used to seed the initial centroids. The experiment were set to 100 maximum of interactions and the value of k varies according to the number of data extracted from the different bibliographic databases. Once the algorithm finishes we have our similar areas based in a bunch of keywords.

5.2. Detect researchers networks

We have discovered similar areas, now it is time to detect what researchers could be interested to work together based in the ares they are working on. We have developed a MapReduce model to accomplish it as you can observe in the figure 2. First, we sorted the name of clusters accord to the unique identifier generated and merged each cluster with their original keyword (before to preprocess). In our final job, we merge the resulting file of the first stage (Sort & Join) and the additional file containing the remaining fields (title, author) from KODAR?s first stage. Finally, we get a file with all the original fields, plus a field showing the cluster that a row belongs.

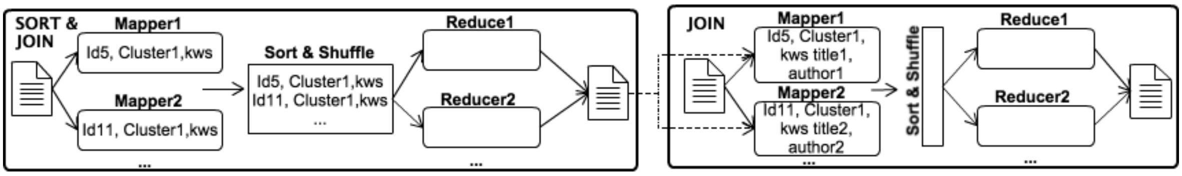

Figura 1. MapReduce model

5.3. Find a general topic area

Our search engine could increase performance in searches by finding a general topic area based in the words that belongs to a cluster. We can respond to specific queries (i.e.: show all

[1] Mahout also use Sequence files to manage input and outputs of MapReduce and store temporary files.

[2] it is used to generate random centroids

researchers working in a specific area or all subareas belonging to a general topic area).

We use WordNet[3] [?] [Smith and Jones 1999] [?] to find synonyms, hypernyms, hyponyms and the concept of a word for all keywords in a cluster. It helps to find a common meaning in the way that words could occur together and find similar meanings. In other words, with the group of word set up we could find a concept or a topic for each cluster.

We applied Collapsed Variational Bayes (CVB) algorithm that is an implementation for Latent Dirichlet analysis (LDA) in Mahout. We use all the words generated by WordNet plus the title and keywords of each publication to find a broader topic based in multiple subtopics described by the keywords. We use Mahout RowId to convert Term Frequency (TF) vectors into a matrix. The CVB algorithm was executed with the following parameters: 1 for the number of latent topics and 20 maximum interactions. This job is applied to each cluster.

Finally the results of three KODAR?s stages are exported in different formats, but one of most import is the Resource Description Framework (RDF) file. Figure 2 shows the concepts and relationships used to export the results. The full arrow symbolize a relationship between classes and the dash arrow symbolize a common relationship.

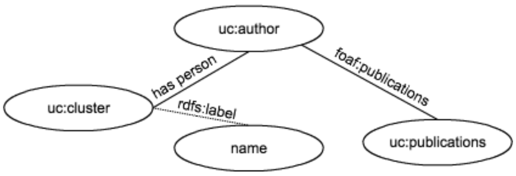

Figura 2. MapReduce model

6. Results

We are going to show the interpretation of a taken sample cluster as result[4]. We find that all the words listed below belongs to the general topic area of physics. Researchers that are working in the areas listed of physics are Fernández Tapia, Jaime E, Torres Arteaga, Christian Alejandro and Aguilar Romero, Gino.At last, a research project could be proposed with people that are working in similar areas.

- Inelastic Scattering
- Flow Measurement
- High Energy
- Fourier Coefficient
- Bose Einstein Correlations
- Monte Carlo
- Three Dimensional
- Center of Mass
- Large Hadron Collider
- Charged Particles
- Correlation Function
- Proton Proton
- Particle Physics
- Experience Repor
- Elliptic Flow
- Heavy Ion Collision
- Particle Production
- Particle Emission

7. Conclusion and Further Work

Lo esencial de esta sección es un resumen de las conclusiones importantes y de sus implicaciones en el área de investigación sobre la que trata el artículo. Tradicionalmente, las conclusiones

[3] it is a lexical database for the English language that is used for text analysis applications.

[4] All results can be analyzed on the web platform: `http://investiguemosjuntos.cedia.org.ec`

ofrecen una descripción (resumida) de los objetivos principales del marco teórico, del rigor metodológico, de los resultados, el uso e impacto de los resultados, la originalidad y el tipo de contribución, y de los desarrollos futuros. [Knuth 1984], [Boulic and Renault 1991] y [Smith and Jones 1999]. Para mayor información sobre el formato de las referencias diríjase al enlace:
`http://diuc.ucuenca.edu.ec/revista-maskana?download=17:guia-autores-maskana`

AGRADECIMIENTOS

En esta sección se agradece de manera cortés por la ayuda: científica, de redacción y técnica (equipo y otros materiales especiales) recibida de cualquier persona o institución. Además, en esta sección se expresa también un reconocimiento por la ayuda financiera externa (como subvenciones, contratos o becas) recibida tanto para la realización de la investigación como para la preparación del artículo. Debe ser breve.

REFERENCIAS

Boulic, R. and Renault, O. (1991). 3d hierarchies for animation. In Magnenat-Thalmann, N. and Thalmann, D., editors, *New Trends in Animation and Visualization*. John Wiley & Sons ltd.

Dirección de Investigación de la Universidad de Cuenca, D. (2014). Directrices para la elaboración de artículos científicos revista maskana de la dirección de investigación de la universidad de cuenca DIUC.

Knuth, D. E. (1984). *The TEX Book*. Addison-Wesley, 15th edition.

Smith, A. and Jones, B. (1999). On the complexity of computing. In Smith-Jones, A. B., editor, *Advances in Computer Science*, pages 555–566. Publishing Press.