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

Research in Iberoamerica has increased in recent years. According to the publication of the State of Science 2015 the number of items registered in the Science Citation Index (SCI) [1] grew by 123 %. Increasing its participation in international databases to increase its local scientific production [2]. One of the most prominent countries is Brazil increased its number of publications by 2.5. However, it has several limitations, as the amount of resources invested in research in contrast to the world average. should be noted that in Latin America is the second fastest growing in the world after Asia [3]. It has a wide range of areas of knowledge, in addition, each country has different strategic ways to address the problems of a region. Which provides a set of solutions that can be an advantage compared to first world countries in the field of research, as these solutions should be able to cope with this heterogeneity in Latin America.

One focus of the IHE (Institutions of Higher Education) of Latin America is to contribute to the sustainable development of society through the cooperation of students and teachers, driven or promoted by research. Currently, certain information about researchers and their bibliographic resources are scattered among various digital repositories or bibliographic databases. When you need to propose projects with several researchers in a specific area belonging to different IHE , raises questions such as: Who works in similar lines of research? or how you can create a network of researchers in a common area when we do not know if they exist? In addition, research to define the profile of a person in analysis, get your articles, in that magazines that were accepted, among others, need to access multiple data sources. Given

that this process is manual, syntactic and different for each source of bibliographic resources available on the 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 Geolink[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. Bibliographical Ontologies.

The Bibliographic Ontology (BIBO) [**?**] developed by Frédérick Giasson and Bruce D'Arcus describe bibliographic things on the semantic Web in RDF. This ontology can be used as a citation ontology, as a document classification ontology, or simply as a way to describe any kind of document in RDF. It has been used to describe publications, citations, conferences, etc. This ontology also helps us to define a common data model for authors and publications on the platform, but it's not enough because we have publications with special properties that don't be in BIBO, so it is necessary to extend this ontology so that our data fits.

An ontology that its similar to BIBO is FaBio the FRBR-aligned [**?**] Bibliographic Ontology that is bears many similarities with BIBO, including its overall scope and intention, and the inclusion of PRISM and DC Terms data properties. this ontology is structured according to the FRBR[*] conceptual model, in which publication entities are described from four different and correlated points of view, those of Work, Expression, Manifestation and Item, each of which is a FRBR Endeavour. FaBiO was developed to describe anything a research scientist might need to reference.

selected ontology to describe the bibliographical resources in the plaform is BIBO because it offers the advantages as semantic reasoning support of an OWL 2 DL, simplicity, expressiveness agil which are not clearly defined in the ontology Fabius. support to make inferences about the data is important if what we do is finding information. Furthermore Fabio offers features to mix with other ontologies which is not necessary in the project at this early stage.

3.2. Apache Marmotta Platform.

Apache Marmotta provide an open implementation of a Linked Data Platform that can be used, extended and deployed easily by organizations who want to publish Linked Data or build custom applications on Linked Data[]. This platform have a recommendation that is being developed by W3C. Among the main features we can mention read-write Linked Data, RDF triple store with transactions, versioning and rule-base reasoning, SPARQL and LDPath query languages, transparent linked data caching, integrated security mechanisms. In December 2013, it has been nominated as ”one of the ASF’s most active projects[] Apache Marmotta comprises some components but specifically has been used JavaEE web application providing the Linked Data server, KiWi a Sesame-based triple store built on top of a relational database and LDClient, a client that allows retrieval of remote legacy resources not available as Linked Data. Through the library Linked client data can consume data from publications from bibliographic databases and whether they are described with some ontologies, or otherwise they are described using ontology BIBO.

3.3. 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.

3.4. K-Means

3.5. Topic Model

3.6. Latent Dirichlet Allocation

4. CONSOLIDATE DATA

Scientific publications of Ecuadorian authors are available in different bibliographics databases on internet at each source varies the information on scientific activity of an author. For example

Scopus sources save affiliation of authors, tables, graphs of publications, authors study areas, etc. DBLP features does not cover. For this reason is necessary make a unification of these bibliographic resources with 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 bibliographics sources, after a data integration process and disambiguation of authors and their publications , and as a third phase defined a method to data update add information to allow a controlled way, facilitating access, discovery and reuse of scientific 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 to describe bibliographic data in a single model.

4.1. Extraction data of bibliographic databases

The extraction process of publications is responsible for obtaining information from scientific articles Ecuadorian authors from various external bibliographic databases 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 available through 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.

4.1.1. Analyse of access on bibliographic resources.

Google Scholar, Scopus, DBLP, Microsoft Academics, Semantic Scholar, GeoLink, etc made available tools that facilitate scientific research. They differ in the form and manner that publish these resources, and the respective policies of each resource or bibliographic database. In some cases it has access to the abstract, references, citations or full article in PDF format and in other cases only the metadata has given publication. As for this platform is considered to use different sources in order to account for the heterogeneity of the sources. Certain data sources have API access to information, but when a source does not have this access can be a problem. The tools that extract should be adapted to the context in which these resources for example websites, sets of documents in pdf format or any type of bibliographic information which is not available through an interface with certain characteristics such as compatibility are published and reliability. If not considered sources like Google Scholar that does not have an access interface information to be analyzed may be limited. Therefore the extraction service adapts to the source in the event that this does not consist of an API, however data from it will not have the same quality as data that come from a source consisting of an API access.

For the development of the first prototype of the project has been selected four bibliographical databases such as Microsoft Academics, Google Scholar, Scopus, DBLP to cover different types of databases and present a scenario in which the main problems involved in making the extraction and integration of bibliographic sources. Subsequently, the platform should allow seamlessly add new data sources and this process should not affect the platform operation.

4.1.2. Analysis of the data models of bibliographic resources data source.

The different bibliographic databases provide their own resources with a logical structure which makes the data model for each source is different despite dealing with the same type of information. Bibliographic resources are not ruled by a standard or comprehensive model encompassing all properties as authors, appointments, conferences, knowledge areas, etc. in the same way. Some features such as DOI, ISBN, format bibliographic references of resources are ruled by International Standard Bibliographic Description (ISBD), ISO 690, etc. However it is not enough if we need a common data model to facilitate the processing of scientific publications.

In Figure [1] you can see the data model between two bibliographic databases: Microsoft Academics and Springer Open Access API API that represent the diversity of models of data between bibliographic databases. This heterogeneity of models represents the challenge of integrating various sources, taking into account that some sources don't publish your data model. Therefore before adding a new data source to the platform must perform an analysis of its structure with respect to models already being used for the purpose of assigning correspondences between model features the new source and model common data.

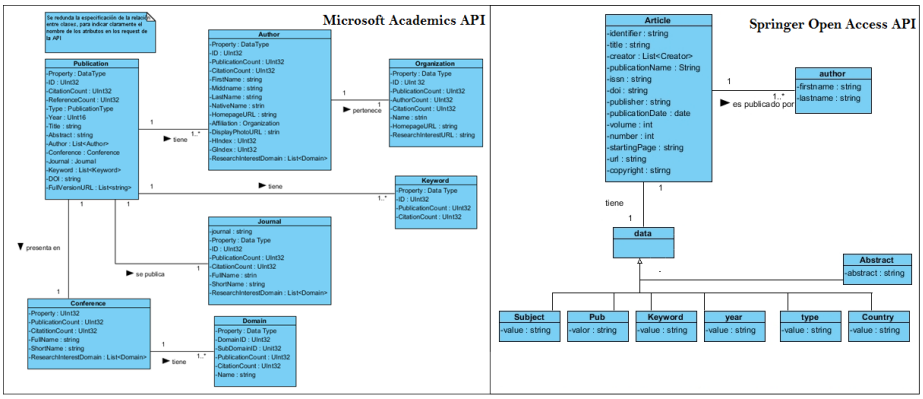

Figura 1. Data Models of Microsoft Academics API and Springer Open Access API.

4.1.3. Architecture extraction service publications.

The architecture for service publications extraction is implemented generically adapted to the source. In Figure * you can see the architecture of extraction service , which have exposed the different data sources to be consulted, which some of them have API access to data. The module requests the service publications extraction publications of an author sending as parameters the author names. The service searches all bibliographical sources external publications of the author and if this are not described in any ontology then describes using BIBO. Finally these data recorded in RDF are sent publications module, then be processed to a central repository.

4.2. integration of authors an publications

The integration process is responsible of unify the publications of all bibliographic sources. The extraction service keeps publications in a respective graph for each source, and it has a graph of publications DBLP, Scopus, etc. Integration service unifies all graphs either publications or authors in a central graph called wkhuska. For this unification we find characteristics between publications and authors in the platform. the origin of the publication is also stored and whether

a publication is similar to another is as described with SameAs property, and equally form the authors, this descriptions are used to disambiguate publications by the same author. As a result the unified data integration process are obtained and available through Apache available marmot SPARQL.

In view of the diversity of data models bibliographical sources a common data model which encompasses the main features of each source publications was raised. The model proposed is described using ontology BIBO (bibliographic ontology), which is an ontology is used to describe bibliographic entities as books, magazines, etc. [10]. This ontology is used to describe bibliographic resources and are listed in RDF. The authors are described using ontology FOAF (Friend of a Friend), it is an ontology used to describe people, their activities and their relationships with other people and objects. The common data model is presented in Figure [2], wherein the different attributes and relationships necessary for the representation of library resources in the platform.

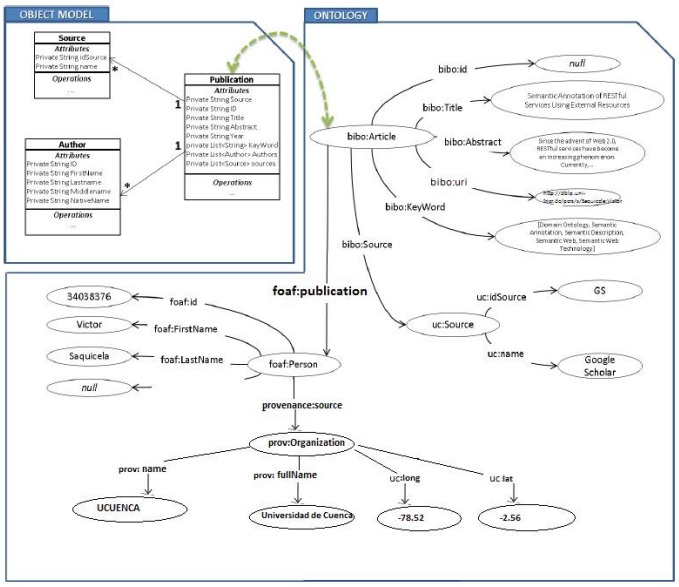

Figura 2. Common data model publications.

The incompatibility between different data models bibliographical databases is handled through a process of mapping, so that a correspondence between the properties of the attributes of publications and bibliographic source model proposed common data set. This process is necessary to specify the time manually correspondences between the properties of publications and model. When adding a new source is must identify the correspondences between data models. However, this process should be done automatically taking into account that this is a defined domain and similar ontologies, (PUTTING THE GEORGE).

It has been necessary to have materialized data authors and publications with the aim of finding correspondences between these locally since the other option is to recover the publications at the time they are needed ie make a request for this to the source in a given moment. The time between making a request to an external source and the mapping takes an average of twenty seconds. So the fact is justified d realize the offerer high availability and speed bibliographic resources to consult the publications of a specific author.

4.3. Disambiguation of authors and bibliographic resources.

The platform have publications from different sources by the same author and to perform this task has several entities of the same person possibly with the same publications, so it is necessary to discover that these authors are the same entity. The prototype developed allows you to define a single record of an author in a central graph based on various entities of the same author of different publications providers. The procedure was performed for this purpose is to analyze the publications of authors who are theoretically the same person with the same publications, unifying them based on a priority set based on experiments. For example the most reliable source is the service Scopus, because other sources of data are inconsistencies when search as the next author ”Juan Pablo Carvallo Vega”and ”Juan Pablo Carvallo Ochoa”if we make this search possible that certain data sources provide us the same publications although locally in our repository are two different people. With the method implemented this problem however is necessary to feed this method more rigorous techniques with the aim of identifying a more concrete form to the authors and their publications present in the platform is. In Figure [3] you can see the diagram of the method implemented for disambiguation of authors and publications. As the first step is loaded into memory publications specific chart such as the graph DBLP or Microsoft Academics, etc. each publication is processed asking if first is already in the repository, if not all the properties of the publication as well as the other publications of the author of the publication being processed at that time we will call P1 is removed. It is also extracted from the central graph in the event of any publication the author of P1 and a comparison between the publications of the graph of the supplier and the publications of the Central graph, if matches between these publications is established is made that the author is already present in the central graph and publication P1 is added to this author. In the event that the author's publications P1 are not in the central graph states that it is the first record of the author and their publications. This process assigns a single central author different resources by the same author repeated in each graph provider publications.

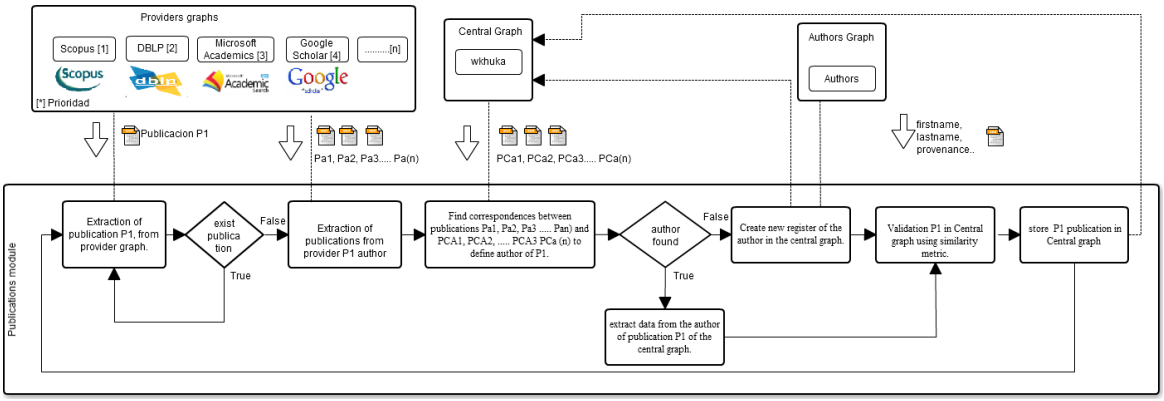

Figura 3. Disambiguation Process.

4.4. Analysis and implementation of data update methods.

Bibliographic resources must be updated at a frequency set according to the number of authors, literature sources, repositories Ecuatorianos authors, etc. The frequency with which a new library resources published author is unknown which is considered an update of information by set of authors. For example it is possible to perform an update of scientific publications of all authors of the University of Cuenca getting their new publications. It is defined methods therefore incremental update not an update of the repository which would take a long time defined by the number of authors being necessary. The process begins with a selection of a set of

authors to update what the authors repository available to each higher education institution such as University of Cuenca (UCUENCA). Army Polytechnic School (ESPE), etc. This process is done with expert intervention activating the selected source manual. Once the set of selected authors to update, the data required for the update is loaded in memory respective scientific publications in a bibliographic database as defined by the expert may be Scopus, DBLP, etc. If changes are published by the authors information is added to the platform.

The number of records of authors on the platform are around ninety-seven thousand so based on experiments when extraction publications all registered authors processing time is too long that can be in is done order of weeks and if a fault occurs in this process should start from scratch. The method implemented timely update has control over the source of publications to use, because such sources as Scopus for example allows a certain value share requests and if this value is exceeded the API blocks access to its resources.

Because they do not have relevant information to more than names and surnames of the authors, it is a difficult task to differentiate between two authors with similar names, so that the process presented presents difficulties arise names like ”Mauricio Espinoza Mejia”and ”Mauricio Espinoza Ana”is a need for more information from the authors to find a relationship between authors and their publications. As a next step in this process is extracting the keyword from registered authors in the source, so that a more concrete relationship between these keywords and the publications of Ecuadorian authors find jobs. The data still contain erroneous information is not yet due to insufficient data with which to disambiguate authors according to their field of study. For example you can consider using keywords of the work done by the authors in order to identify whether the publications obtained by the service relate to the keywords of local works of the author.

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.

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

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 5. 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.

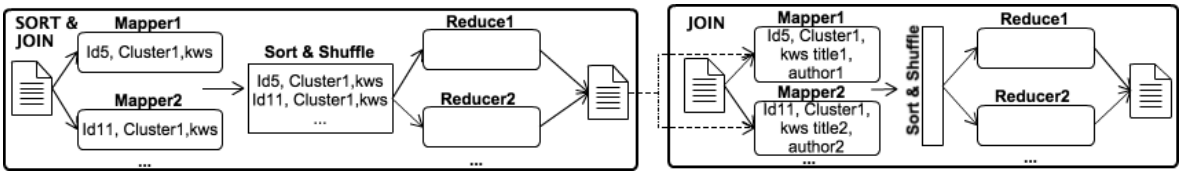

Figura 4. MapReduce model

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 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.

[2] it is used to generate random centroids

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

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 5 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.

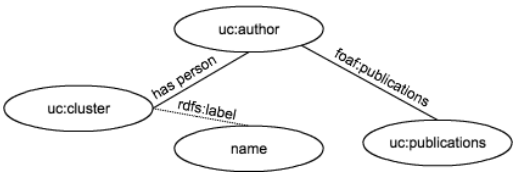

Figura 5. MapReduce model

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 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`

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

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.