Universidad Politécnica de Madrid

Master Thesis

# RESSIST: A Research Object-based Recommender System

*Author:*
Carlos Badenes Olmedo

*Supervisor:*
Oscar Corcho García

*A thesis submitted in fulfilment of the requirements*
*for the degree of Master of Science*

*in the*

Ontology Engineering Group
Artificial Intelligence

July 2015

UNIVERSIDAD POLITÉCNICA DE MADRID

# *Abstract*

ETSI-Inf
Artificial Intelligence

Master of Science

**RESSIST: A Research Object-based Recommender System**

by Carlos Badenes Olmedo

The scientific method is a methodological approach to the process of inquiry – in which empirically grounded theory of nature is constructed and verified [14]. It is a hard, exhaustive and dedicated multi-stage procedure that a researcher must perform to achieve valuable knowledge. Trying to help researchers during this process, a recommender system, intended as a researcher assistant, is designed to provide them useful tools and information for each stage of the procedure. A new similarity measure between research objects and a representational model, based on domain spaces, to handle them in different levels are created as well as a system to build them from OAI-PMH (and RSS) resources. It tries to represents a sound balance between scientific insight into individual scientific creative processes and technical implementation using innovative technologies in information extraction, document summarization and semantic analysis at a large scale.

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **DCMI** | **D**ublin **C**ore **M**etadata **I**nnovation |
| **JSD** | **J**ensen **S**hannon **D**ivergence |
| **LDA** | **L**atent **D**irichlet **A**llocation |
| **LSA** | **L**atent **S**emantic **A**nalysis |
| **MOEA** | **M**ulti-**O**bjective **E**volutionary **A**lgorithm |
| **PLSA** | **P**robabilistic **L**atent **S**emantic **A**nalysis |
| **PLSI** | **P**robabilistic **L**atent **S**emantic **I**ndexing |
| **OAI-PMH** | **O**pen **A**rchive **I**nitiative **P**rotocol for **M**etadata **H**arvesting |
| **RDF** | **R**esource **D**escription **F**ramework |
| **RO** | **R**esearch **O**bject |
| **TF-IDF** | **T**erm **F**requency - **I**nverse **D**ocument **F**requency |
| **URI** | **U**niform **R**esource **I**dentifier |
| **URL** | **U**niform **R**esource **L**ocator |
| **VSM** | **V**ector **S**pace **M**odel |

# Chapter 1

# Introduction

The scientific method is probably the most widely used method or way of knowing the unknowns. It is a methodological approach to the process of inquiry – in which empirically grounded theory of nature is constructed and verified [14] that increases the human knowledge based on systematic observation, classification and interpretation. In addition, it is characterized by objectivity, generality, verifiability and creditability to ensure an unbiased, general and impersonal study [63]. This process consists of a series of steps or actions that are important to execute a specific research in an effective way, such as: (1) *define a research problem*, (2) *review literature*, (3) *write a hypothesis*, (4) *design the research*, (5) *collect data*, (6) *analyze the collected data* and (7) *interpret results and report* [66].

While each of these steps is important, the first one is crucial because the whole study is designed around this defined goal. It should address a unique issue, building upon previous research and scientifically accepted fundamentals. It has several considerations:

- **Interest**: It should be interesting for the researcher and for the research community.
- **Magnitude**: It should be manageable, specific and clear to have enough time and resources to solve it .
- **Expertise**: It requires an adequate level of expertise for carrying out the research.
- **Relevance**: It should contribute to answering questions of quantified importance to the end users of research [68].

- **Availability of data**: It should make sure that data is available.

Taking into account these considerations, there are several steps to formulate a *research problem*: First of all, identify a broad field or subject area of interest and split it into subareas, then select the most interesting one formulating research questions and, finally, assess the objectives.

Once a ***research problem*** is identified, the researcher has to contextualize it. In doing so, an initial ***literature review*** is needed to build a theoretical and conceptual framework using existing literature in the research area of study.

When the problem is clearly identified, it is time to *write a set of hypotheses* that need to be proved with new experiments and observations. Usually it is the result of a process of inductive reasoning from observations to create a testable, falsifiable and realistic statement. It is a suggested explanation of a phenomenon, so it should be testable, taking into account current knowledge and techniques, realistic and also verifiable, by statistical or analytical means, to allow a verification or falsification. It eventually becomes a theory, even then it can still be falsified or adapted.

Now, the researcher begins to ***design the research*** defining variables according to what will be measured. *Independent variables* are those that the researcher would like to measure (the cause), *dependent variables* represent the effect and *extraneous variables* are a particular case of *independent variables* that are not related to the purpose of the study, but may affect the *dependent variable*.

In order to make accurate measurements, a ***collection of data*** about the research problem must be gathered. There are two types of data: *primary data*, generally accepted as original data that can be collected through questionnaires, schedules, interviews, sensoring and so on; and *secondary data*, also known as *published data*, which are not originally collected but rather obtained from published sources such as publications in books, magazines, reports, etc.

In this stage of research, the collected data should be processed and analyzed. ***Analysis of collected data*** represents the way of testing hypotheses and support the approach of achievement of findings and so the conclusion of the research. This process may be manual or automatic. The researcher must provide details about methods, procedures and models used during this process.

Finally, the researcher has to present the ***interpretation of results and the report*** derived from this study in a structured and logical manner following a systematic, chronological or psychological order. The most important thing is to prune out irrelevant information and findings.

As shown above, the research process is a hard, exhaustive and dedicated multi-stage procedure that researchers perform to achieve valuable knowledge. Each of its stages includes difficulties and challenges that may block the entire process. Trying to help researchers during this procedure, a recommender system, intended as a researcher assistant, is proposed to provide them useful information for each stage. It represents a sound balance between scientific insight into individual scientific creative processes and technical implementation using innovative technologies in information extraction, document summarization and semantics. This study comes under the Dr Inventor project, a European project built on the vision that technologies have great potential to supplement human ingenuity in science by overcoming the limitations that people suffer in pursuing scientific discovery.

This document provides an explanation of the theoretical background used to make predictions and to infer knowledge from scientific researches devoted to data gathering techniques from content providers. Regarding analysis of research projects, we show a revision of the state of the art focused on those works that attempt to measure similarity between textual resources and we describe how we used these techniques for measuring the similarity between *Research Objects* (RO). In a similar manner, concerning recommender systems, we introduce past related work to provide an overview of these type of systems using semantic and non-semantic information to make predictions. After that, we introduce our system and more details about several functions that it offers to researchers based on semantic analysis and machine learning techniques.

The document is organized in the following way: Section 2 reviews the state of the art in semantic and non-semantic-based recommender systems. It also presents *Research Objects* and the Dublin-Core annotation, as well as topic models and why they are useful to summarize data. We introduce the *Open Archive Initative Protocol for Metadata Harvesting (OAI-PMH)* and some statistics gathered from content providers implementing it. Section 3 presents in detail what are the problems that we are trying to solve with respect to how we can provide better support to the research process. In Section 4 we

show some statistics about OAI-PMH data providers and introduce our harvester client. Section 5 presents how we have built an LDA model using a genetic algorithm so as to describe all researches in a more accurate way. Section 6 details how research objects are internally described to gather meaning and content information together. In Section 7 we detail the elements that compose our domain, how they are organized and what is the similarity metric created to measure how *Research Objects* are connected. Section 8 describes some recommendations that the system could offers to promote scientific creativity and help researchers during the research process. To conclude, suggestions for future work and conclusions are proposed in Section 9.

# Chapter 2

# Background and State-of-the-Art

## 2.1 Research Objects

Reproducible science has become a field of research in its own right [49]. The reproducibility of a scientific study depends on the careful description of the original experiment, including the methods and tools used to perform the experiment, the substrate on which to perform the experiment, and the precise experimental setup, including all necessary influences from the environment. When the result is meant to be present after post-processing, it is also imperative to provide the details of the processing steps [13].

The useful outcomes of research are not just traditional publications. Instead they are everything else that goes into, and supports an investigation. A research life cycle includes steps which need to be documented and described in order to be reproduced. Several guidelines have been developed about it such as the *Minimum Information Requested in the Annotation of Models* (MIRIAM) [46] and the *Minimum Information About Simulation Experiments* (MIASE) [71]. The outcome is a findable, accesible, interoperable and reusable resource that adds value to the research.

Our system must be able to understand all this information to learn as much as possible about the inner process related to the research and make valuable inferences based on similarities between them. In that regard, the *Research Object* model allow us to handle researches as machine-readable resources being easier sharing and/or exchanging information.

Research Objects are semantically rich aggregations of resources that bring together data, methods and people in scientific investigations. Their goal is to create a class of artifacts that can encapsulate the digital knowledge associated to an investigation and provide a mechanism for sharing and discovering assets of reusable research and scientific knowledge. It is a combination of *aggregation* (reusing *Object Reuse and Exchange (ORE)*), *annotation* (reusing the *Annotation Ontology*) and RO *ontologies* such as *ro* (core structure), *wf-desc* (workflows) and *wf-prov* (provenance).

### 2.1.1 Benefits

***Unique identifiers*** are used as names for things (e.g. DOIs for publications or ORCIDs for researchers) with the objective of avoiding ambiguity about resources (publications, researches) and ensuring that any resource is easier to be found.

Elements that are related or part of a broader investigation or study can be ***aggregated*** as artefacts that make the research potentially useful to someone else because it can be referred to or ***cited as a whole***. Also ***metadata information*** about how they relate to each other, or where they came from, or what are the rights, or any other information is provided along with the resource.

### 2.1.2 Distribution

Recently, the Wf4Ever project developed the *Research Object Bundle (RO Bundle)* [74]. It is a ZIP file containing a manifest, annotations and some or all of its aggregated resources for the purposes of exporting, archiving, publishing and transferring research objects. Its structure is based on the *Adobe Universal Container Format (UCF)*.

The advantage of encapsulating everything in a single file is to know about a specific project, including the instructions on how to handle the archive and interpret it. Similar examples from the domain of computer science would be the packages of the Java Archives, the Open Document Formats or recently the *COMBINE Archives* [13], that enable the exchange of all information required to reproduce a modelling project encoded using the *Open Modelling EXchange format (OMEX)* and presented as a ZIP file. Its default file extension is *.omex* and additional extensions are available to indicate what

is the main standard format used within the archive. This helps users choose between different archives, and select appropriate software tools with which to open them.

## 2.2 OAI-PMH

Modern researchers have access to large archives of scientific articles. These archives are growing as new articles are placed online and old articles are scanned and indexed. Our system is based on *research objects* so we need to collect them from online services. As far as we know, *research objects* are not currently available from any web service, so we need to create them from existing publishing services.

The *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH) [53] is a client-server protocol for exchanging data in numerous formats which is the *de facto* standard for metadata sharing between digital libraries in distributed environments [45]. It is based on two roles: *data providers*, as storage systems that support OAI-PMH as a means of exposing metadata, and *service providers*, as clients that use metadata harvested via the OAI-PMH as a basis for building value-added services.

We can create *research objects* combining both the metadata and the referenced resource published by these providers. In such a way our system will play the role of *service provider* operating a *harvester*, i.e client application that issues OAI-PMH requests, to collect metadata from online *data providers*.

In this domain, three distinct entities appear related to the metadata [53]: *resource*, *item* and *record*. A *resource* is the object that metadata is "about" and any other information about it is outside the scope of OAI-PMH. An *item* is a container that stores or dynamically generates metadata about a single resource in multiple formats, each of which can be harvested as a *record*. A *record* is metadata in a specific format such as an XML encoded byte stream, in response to a protocol request.

### 2.2.1 Requests and Responses

The requests, or *verbs*, defined in the OAI-PMH protocol are the following [53]:

- ***GetRecord***: retrieve an individual metadata record from a repository by *identifier*, i.e unique id, and *metadataPrefix*, i.e. metadata format, such as *oai_dc*, *olac*, *perseus*, *oai_marc* and so on.
- ***Identify***: retrieve information about a repository such as *name*, *base url*, *protocol version*, *earliest datestamp*, *delete record*, *granularity*, *admin email*, *compression* and/or *description*.
- ***ListIdentifiers***: abbreviated form of *ListRecords*, retrieving only *headers* rather than *records* by a *temporal window*, *metadata prefix*, *set* (selecting criteria) and *resumption token* to be able to resume an incomplete list.
- ***ListMetadataFormats***: retrieve the metadata formats available from a repository.
- ***ListRecords***: harvest records from a repository by a *temporal window*, *metadata prefix*, *set* (selecting criteria) and *resumption token* to be able to resume an incomplete list.
- ***ListSets***: retrieve the set structure of a repository, i.e. list of fields to make selecting criteria.

### 2.2.2 Dublin Core Annotation

As mentioned before, a data provider can use several metadata formats. The format used in our harvester client was *Dublin Core Metadata Innovation* (DCMI) [24], which is associated with the reserved *metadataPrefix oai_dc* in OAI-PMH. DCMI is one of the most popular vocabularies for use with *Resource Description Framework* (RDF). The *DCMI Abstract Model* was designed to bridge the modern paradigm of unbounded, *Linked Data* graphs with the more familiar paradigm of validatable metadata records like those used in OAI-PMH.

It provides an agreement for the development of interoperable online metadata standards for a broad range of purposes and of business models. Four *levels of interoperability* are presented for determining the scope of a project that wants to be *"Dublin Core-compatible"* and to set expectations for users of *"Dublin Core-compatible"* specifications [24]: (1) *Shared Term Definitions*, (2) *Formal Semantic Interoperability*, (3) *Description*

*Set syntactic interoperability* and (4) *Description Set Profile interoperability*. Among these values, the first one, *Shared Term Definitions* which establishes interoperability among metadata-using applications based on shared natural-language definitions, is the best to describe how *data providers* and *service providers* operate in OAI-PMH since they agree what terms to use in their metadata and how those terms are defined.

According to the DCMI-based XML schema used in OAI-PMH (*Listing* A.1), a resource may be described by *title*, *creator*, *subject*, *description*, *publisher*, *contributor*, *date*, *type*, *format*, *identifier*, *source*, *language*, *relation*, *coverage* and *rights*.

These metadata terms are maintained and specified by DCMI as follows [23]:

- *title*: a name given to the resource.
- *creator*: an entity primarily responsible for making the resource.
- *subject*: the topic of the resource.
- *description*: an account of the resource which may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource.
- *publisher*: an entity responsible for making the resource available.
- *contributor*: an entity responsible for making contributions to the resource.
- *date*: a point of period of time associated with an event in the lifecycle of the resource.
- *type*: the nature or genre of the resource. Usually described by *DCMI Type Vocabulary* (DCMITYPE).
- *format*: the file format, physical medium or dimensions of the resource.
- *identifier*: an unambiguous reference to the resource within a given context. Recommended best practice is to identify the resource by means of a string conforming to a formal identification system.
- *source*: a related resource from which the described resource is derived.
- *language*: a language of the resource. Recommended best practice is to use a controlled vocabulary such as RFC 4646.

- *relation*: a related resource. Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.
- *coverage*: the spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
- *rights*: information about rights held in and over the resource.

## 2.3 Topic Models

Since we want to build a *research object*-based recommender system, we need to define how ROs will be featured. We are looking for a model that describes resources not only by individual features (e.g. term frequencies), but also by features shared with others (e.g. subjects) in an adequate way to be used in large data sets, so first, we list the advantages of using topic models to represent this type of resources, after that we present how it works compared to other approaches such as *tf-idf* and, finally, we show why we have used this model.

The utility of topic models stems from the property that the inferred hidden structure that they reveal resembles the thematic structure of the collection. A *hidden structure* is a topic structure such as *topics distribution*, *per-resource topic distributions* or *per-resource per-word topic assignments*. In turn, a *topic* is a distribution over terms that is biased around those associated under a single theme. This interpretable *hidden structure* annotates each resource in the collection and these annotations can be used to deeper analysis about relationships between resources. In this way, topic modeling provides us an algorithmic solution to managing, organizing and annotating large collections of *research objects* according to their *topics*, i.e according to the distribution of their terms in *topics*.

Topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over the time [15]. They do not require any prior annotations or labelling of the documents. The topics emerge, as hidden structures, from the analysis of the original texts. The main challenge is how to use the observed resources to infer a hidden topic structure. It could be seen as *reversing* the generative process, i.e. what is the hidden structure that likely generated the observed collection?

### 2.3.1 Automatic Summarization

Popular algorithms reduce each resource in the corpus to a vector of real numbers taking into account frequencies of words, e.g. The *tf-idf* scheme takes into account the number of occurrences of each word compared to an inverse document frequency count measuring the number of occurrences of a word in the entire corpus. Though they present some appealing features such as identification of discriminative words, they are unable to reveal inter or intra resource statistical structure.

Other approaches capture most of the variance in the collection and even some aspects of basic linguistic notions, e.g. synonymy and polysemy, such as *latent semantic analysis (LSA)* but they cannot express more complex relationships between documents, between words and between documents and words. Since they are *discriminative models*, they provide a model only for the target variable, documents, conditional on the observed variables, words, ignoring hidden structures such as *topics*.

Thus, to describe these hidden structures, i.e. these complex relationships, not only a topic model algorithm is required, but also a *generative* topic model algorithm such as *Latent Dirichlet Allocation* (LDA). This will enable to organize and summarize *research objects* at scale what would be impossible by any other manner, preserving the essential statistical relationships that are useful for tasks such as classification, summarization and similarity and relevance judgements.

### 2.3.2 Latent Dirichlet Allocation

The simplest generative topic model is *latent Dirichlet allocation* (LDA) [16]. This and other topic models such as *Probabilistic Latent Semantic Analysis (PLSA)* [39] are part of the larger field of *probabilistic modeling*. They are well-known latent variable models for high dimensional count data, such as text data in the *bag-of-words* representation or any other count-based data representation but, while LDA has roots in LSA and PLSA (it was proposed as a generalization of PLSA), it was cast within the generative Bayesian framework to avoid some of the overfitting issues that were observed with PLSA. As mentioned before, since PLSA is a *discriminative model*, it is unable to describe topics, i.e. hidden structures, but LDA built a generative model to avoid that limitation.

In generative probabilistic modeling, data is treated as arising from a generative process that includes *hidden variables*. This generative process defines a *joint probability distribution* over both the observed ($O$) and hidden random variables ($\mu$). Then data is analyzed by using that joint distribution to compute the *conditional distribution* of the hidden variables given the observed variables $p(\mu \mid O)$. This conditional distribution is also called the *posterior distribution*. In LDA, the observed variables are the words of the documents, the hidden variables are the topic structure and the generative process is the problem of computing the posterior distribution, i.e. the conditional distribution of the hidden variables given the documents:

$$p(O, \mu) = p(O \mid \mu) \cdot p(\mu) = p(\mu \mid O) \cdot p(O) \tag{2.1}$$

This statistical model tries to capture the intuition that documents exhibit multiple topics. Each document exhibits the topic in different proportion, each word in each document is drawn from one of the topics, where the selected topic is chosen from the per-document distribution over topics. All the documents in the collection share the same set of topics, but each document exhibits these topics in different proportion. Documents are each represented as a vector of counts with $W$ components, where $W$ is the number of words in the vocabulary. Each document in the corpus is modelled as a mixture over $K$ topics, and each topic $k$ is a distribution over the vocabulary of $W$ words. Each topic is drawn from a Dirichlet with parameter $\beta$, while each document's mixture is sampled from a Dirichlet with parameter $\alpha$. Formally, a *topic* is a multinomial distribution over words of a fixed vocabulary representing some concept.

The Dirichlet distribution is a continuous multivariate probability distribution parameterized by a vector of positive reals whose elements sum to 1. It is *continuous* because the relative likelihood for a random variable to take on a given value is described by a probability density function, and also it is *multivariate* because it has a list of variables each of whose value is unknown. In fact, the Dirichlet distribution is the conjugate prior of the categorical distribution and multinomial distribution.

From a collection of documents, LDA infers: *per-word* topic assignment, *per-document* topic proportions and *per-corpus* topic distributions. Exact inference, i.e. computing the posterior over the hidden variables, for this model is intractable [16], then a variety

of approximate algorithms have been proposed [7] such as *collapsed Gibbs sampling (CGS)*, *variational Bayesian inference (VB)*, *collapse variational Bayesian inference (CVB)*, *maximum likelihood estimation (ML)* and *maximum a posteriori (MAP)*.

Unlike a clustering model, where each document is assigned to one cluster, LDA allows documents to exhibit multiple topics. For example, LDA can capture that one article might be about "biology" and "statistics", while another might be about "biology" and "physics". Since LDA is unsupervised, the themes of "physics", "biology" and "statistics" can be discovered from the corpus; the mixed-membership assumptions lead to sharper estimates of word cooccurrence patterns.

## 2.4 Recommender Systems

There has been much work done on developing new approaches for recommendation systems over the last decade. The interest in the area still remains high because personalized recommendations have many practical applications.

The two main elements in a recommender system are *users* and *items*. Depending on how they are processed, recommender systems are typically grouped into two categories: *content-based filtering* and *collaborative filtering* (*Figure* 2.1).

### 2.4.1 Content-based Filtering

The *content-based filtering* approach creates a profile for each user or item to characterize its nature. The profiles allow programs to associate users with matching items, i.e analyze the items to extract attributes/features from them and recommend items with similar attributes to an item the user likes. The adoption of a *content-based* paradigm has several advantages when compared to the *collaborative* approach [69]: *user independence*, i.e these systems exploit solely ratings provided by the active user to build her own profile, *transparency*, i.e explanations on how the recommender system works can be provided by explicitly listing content features or descriptions that caused an item to occur in the list of recommendations, and *new item*, i.e. they are capable of recommending items not yet rated by any user. However, it has several shortcomings: *limited content analysis*, i.e it depends on domain knowledge and has a natural limit in the number and

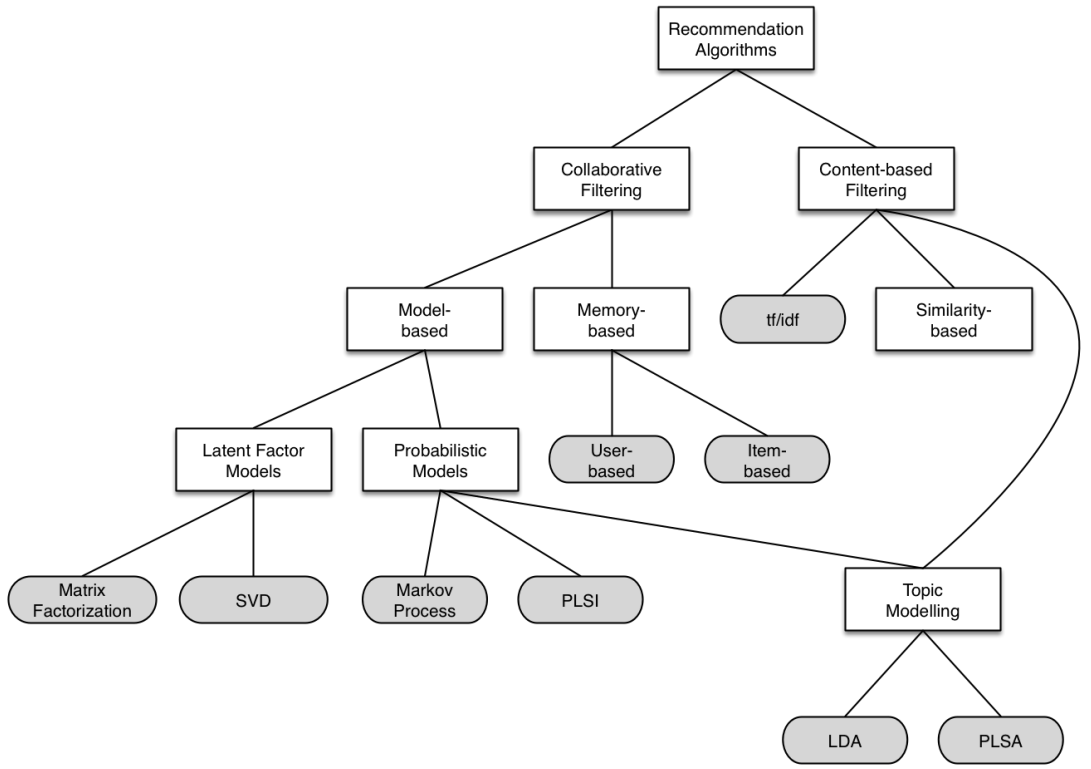



FIGURE 2.1: A hierarchy of recommender systems

type of features that are associated with the items they recommend, *over-specialization*, i.e user is going to be recommended items similar to those already rated so it has no inherent method for finding something unexpected, and *new user*, i.e enough ratings have to be collected before this system can really understand user preferences and provide accurate recommendations.

### 2.4.2 Collaborative Filtering

The alternative, *collaborative filtering*, recommends items to a user based on other users with similar patterns of selected items or on past user behaviour, for example previous transactions or items ratings, without requiring the creation of explicit profiles. This approach analyzes relationships between users and interdependencies among items to identify new user-item association. It is domain free and can address data aspects that are often elusive and difficult to profile using *content-based filtering* but it suffers from what is called the *cold-start* problem [44], due to its inability to address the system's new items and users, i.e it performs poorly if there are no sufficient numbers of co-rated items in a given rating data.

The two primary areas of *collaborative filtering* are the *memory-based* methods, also known as *neighbourhood-based*, [17],[29] and *model-based* methods. *Memory-based* methods are centered on computing the relationships between items, *item-based* approaches [30], or alternatively, between users, *user-based* approaches [35] [60] [50]. These approaches evaluate a user's preference for an item based on ratings of *neighbouring* items by the same user. A item's neighbours are other items that tend to get similar ratings when rated by the same user. In short, an item might be interesting to an active user if the item is appreciated by a set of similar users (neighbours) or she/he has appreciated similar items in the system. So these methods use a similarity measure for finding similar users to an active user or similar items on which she/he rated such as *Pearson Correlation Coefficient*, *Cosine similarity* or *Jaccard Index*. But, since they utilize ratings of only co-rated items when computing similarity between a pair of users or items, their similarity measures are not suitable in a sparse data [75], although recently a new measure has been proposed to prevent this problem [57].

However, this type of methods is not adequate for us because it is based on *explicit feedback*, and we would like to create a system based only on items, i.e. *research objects* and all additional information derived from them and from users, *implicit feedback*. Two different techniques can be adopted for recording user's feedback. When a system requires the user to explicitly evaluate items, this technique is usually referred to as *explicit feedback*. The other technique, *implicit feedback*, does not require any active user involvement, in the sense that feedback is derived from monitoring and analyzing user's activities. In general, approaches based on *explicit feedback* are not useful for us because they limit researchers to specific areas. For instance, a statistician may miss a relevant paper in economics or biology because the two literatures rarely cite, i.e rate, each other; and she/he may miss a relevant paper in statistics because it was also missed by the authors of the papers that she/he has read. In our opinion, one of the main opportunities of recommendation systems in the scope of research publications is to inform researchers about literature that they might not be aware of.

*Model-based* algorithms build models to describe the behaviours of users using training data and utilize the trained models to predict the users' preference on the items unseen in the training data. The main advantage of this approach is that it does not need to access the whole set of data once the model is built. Examples of this approach

include the *latent factor models* [1] [2], *probabilistic models* [34] [58] and combined ones of *probabilistic and latent factor models*.

*Latent factor models* are an alternative approach that tries to explain the ratings by characterizing both items and users on a number of factors inferred from the rating patterns. The key idea behind this is to project the users and items into a smaller dimensional space, such lower dimensional projections are called *factors*, thereby clustering similar users and items. Thus, the interest, *similarity*, of a user to an unrated item is measured and the most similar item/s is recommended to the user. Some of the most successful realizations of *latent factor models* are based on *matrix factorization* [44] [77]. In its basic form, *matrix factorization* characterizes both items and users by vectors of factors inferred from item rating patterns [64]. High correspondence between item and user factors leads to a recommendation. A user's predicted rating for an item, relative to the item's average rating, would equal the dot product of the item's and user's factor vectors. These methods have become popular in recent years due to good scalability with predictive accuracy. Since *latent factor models* can take advantage of not only the explicit relationships between users but also the implicit relationships between two users without any connection, they are effective for recommending users in social networks where people form different clusters according to their preferences. In this case, all of such methods estimate the *latent factor vectors* of users where each factor measures the extent to which the user is connected with the other users according to its corresponding factor. However, they do not take into account the auxiliary information typically associated with items and, as mentioned before, since they use only relationships but not the contents generated by users, the performance suffers when a user does not have enough number of relationships yet, *cold-start*. In addition, since this approach depends on item rating and likely users rate only a small percentage of possible items, the quality of results is limited.

More complex *probabilistic models* were later proposed in various recommendation applications. The *Markov process* was used to model the purchasing process of market basket data. Recommendation systems using the *probabilistic latent semantic indexing* (PLSI) were also developed such as TWITOBI [42] that uses a probabilistic model of the behaviour of writing tweet messages by generalizing the PLSI. This model assumes that the topics are not only selected by a user but also chosen under the influence of the users whom the user follows. However, this model does not capture the generative process of

establishing friend relationships. These algorithms build models based on the relationships between users and items, but not based on more complex relationships between hidden elements. Furthermore, PLSI suffers from the overfitting problem compared to other probabilistic models such as LDA [19].

Thus, LDA [16] appears as a generative algorithm that models both latent topics and hidden communities of users. The initial algorithm, TWITOBI, was then updated to use it in TWILITE [43] which recommends top-K users to follow and top-K tweets to read for a user along with a *matrix factorization* to connect friends. Nevertheless, these techniques not consider the relationships between users and we cannot use them directly in our system where *research objects*, that are created by authors, would be connected to others based on common authors.

As mentioned before, in the scope of research publications, historically one way that researchers find articles is by following citations in other articles that they are interested in. This is an effective practice but it limits researchers to specific citation communities and it is biased towards heavily cited papers. A complementary method of finding articles is keyword search. This is a powerful approach, but it is also limited. Forming queries for finding new scientific articles can be difficult as a researcher may not know what to look for; search is mainly based on content and is only good for directed exploration, while many researchers would also like a "feed" of new and interesting articles.

Recently, a machine learning algorithm called *Collaborative Topic Regression* (CTR) [72] was developed for recommending scientific articles to users of online archives based on traditional *collaborative filtering* and *topic modeling*. This algorithm uses the other users' libraries and the content of the articles to form its recommendations. For each user, the algorithm can find both older papers that are important to other similar users and newly written papers whose content reflects the user's specific interests. This approach combines ideas from *collaborative filtering* based on *latent factor models*, i.e. *matrix factorization*, and content analysis based on *probabilistic topic modeling*, i.e. *LDA*. Like *latent factor models*, this algorithm uses information from other users' libraries. For a particular user, it can recommend articles from other users who liked similar articles. *Latent factor models* work well for recommending known articles, but cannot generalize to previously unseen articles. To generalize to unseen articles, the algorithm uses *topic modeling*. Topic modeling provides a representation of the articles in terms of latent

themes discovered from the collection. So, this component can recommend articles that have similar content to other articles that a user likes. The topic representation of articles allows the algorithm to make meaningful recommendations about articles before anyone has rated them. The algorithm naturally balances the influence of the content of the articles and the libraries of the other users. An article that has not been seen by many will be recommended based more on its content and an article that has been widely seen will be recommended based more on the other users. This method provides better performance than *matrix factorization* methods alone, indicating that content can improve recommendation systems. Further, while traditional collaborative filtering cannot suggest articles before anyone has rated them, this method can use the content of new articles to make predictions about who will like them. However the system depends on users' opinion instead of a more complex users' profile based on her/his publications.

That approach as well as hybrid recommendation models that combine collaborative and content information [36] [18] are a good starting point for our system, where we will try to create user profiles (*authors*) without directly requiring information from them such as users' libraries or user ratings but also considering derived information such as their publications and/or their research topics. Our system is a research object-based recommender system, so we should use all useful knowledge about existing keyword-based recommender systems as previously mentioned and create what we need to handle *research objects* as items and *authors* as users, as well as define new functionalities to convert this system into a helpful research assistant for the research process.

# Chapter 3

# Objectives

## 3.1 General Objective

As mentioned in Section 1, the research process is a multi-stage procedure that includes enough challenges and difficulties to block the entire process in any of its stages. Trying to avoid obstructions, we want to create a system using the current technologies that collects and analyzes ROs at a large scale, and then extracts useful knowledge from them to suggest helpful information to researchers during their research process, even at the beginning when there is no specific research problem defined yet.

The system should offer computational creativity to users who are practising scientists. It is oriented towards helping research scientists by discovering analogical comparisons between academic documents and related sources for their consideration. In this way, the system acts as a *creativity assistant*, while its cognitively inspired architecture also offers one possible model of people thinking creatively.

The final aim is to create a researchers' personal research assistant by reporting on a wide variety of relevant concepts through machine-powered search and visualization. In addition it will also be an indexer of *Research Objects* where anyone can search by *context-based criteria* such as title, authors, license rights, etc and/or *content-based criteria* such as topics, similarity, etc.

## 3.2 Specific Goals

Taking into account the overall objective, i.e *extracting knowledge from research objects to promote scientific creativity in researchers*, some specific goals must be defined to establish a research line that allow us to reach that general objective.

First of all, to create a large and varied *research object dataset*:

- **G01**: ***Create a research object-based dataset***
  - **G01.1**: *Download research resources*
  - **G01.2**: *Convert the resources to research objects*
  - **G01.3**: *Store the research objects*

Then, to compare the research objects:

- **G02**: ***Create a model that describes, as accurately as possible, the research objects***
  - **G02.1**: *Analyze the differences between descriptive, discriminative and generative models and select one of them*
  - **G02.2**: *Create a representation of research objects for the selected model*
  - **G02.3**: *Prepare an automatic procedure to increase the accuracy of the model*
- **G03**: ***Define a similarity measure between research objects based on that model***
  - **G03.1**: *Define the most representative parts of a research object*
  - **G03.2**: *Create a similarity measure based on these parts*
- **G04**: ***Compare all research objects using the similarity measure***
  - **G04.1**: *Implement the similarity measure in a scalable way*
  - **G04.2**: *Create a matrix with the similarity measures of all research objects*

Finally, to make recommendations to provide better support to the research process:

- **G05**: ***Suggest research objects***

- **G06**: ***Suggest research topics***
- **G07**: ***Suggest collaborations***

## 3.3 Contribution of the Thesis

Trying to reach these goals, we have developed the following contributions:

- **Contribution 1:** ***"OAI-PMH/RSS Hoarder and Harvester"***: It responds to the specific goals: *G01.1, G01.2* and *G01.3*.
- **Contribution 2:** ***"LDA configuration using evolutionary multi-objective optimization"***: It responds to the specific goal: *G02.3*.
- **Contribution 3:** ***"Research Object Representation"***: It responds to the specific goals: *G02.1, G02.2*.
- **Contribution 4:** ***"Research Object Similarity Measure"***: It responds to the specific goals: *G03.1, G03.2, G04.1* and *G04.2*.
- **Contribution 5:** ***"Design of Research Object-based recommendations"***: It responds to the specific goals: *G05, G06* and *G07*.

# Chapter 4

# Contribution 1: OAI-PMH/RSS Hoarder and Harvester

As mentioned in Section 2, we have developed a hoarder application [11] and a harvester application [8] that issues OAI-PMH [53] requests to providers to collect both metadata and resources and build *research objects* from them. In short, for each record listed in the OAI-PMH Data Repository (or RSS site), the OAI-PMH/RSS Hoarder creates one metadata file based on the DC-Terms used to describe the resource and downloads the related content, usually as a *pdf* or a *doc* file for publications or as a *txt* when an external file does not exist, e.g. rss news. After that, the Harvester application combines both informations to create a data structure based on the research object model to be managed by the RESSIST system (Figure 4.1).

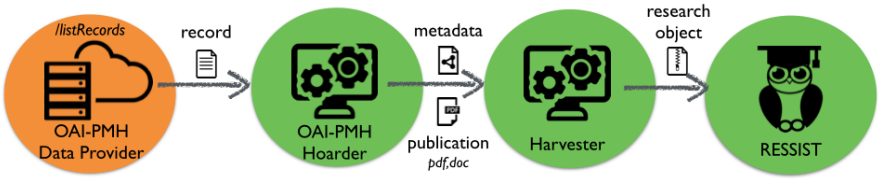
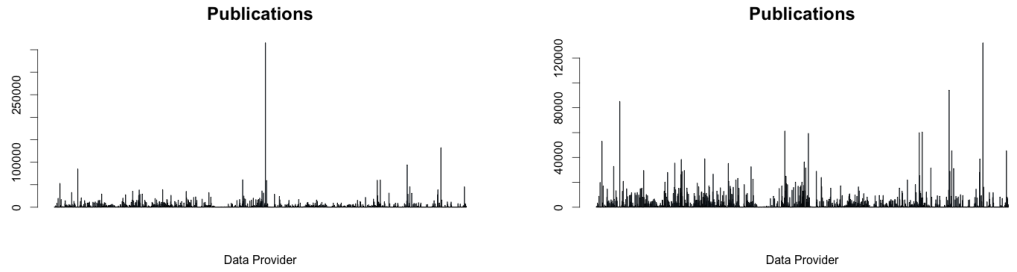



FIGURE 4.1: Building process of Research Objects from OAI-PMH resources

## 4.1 DCMI Terms Usage

Before creating these applications, we developed a lightweight hoarder application [9] that checks metadata (no resource download) from repositories listed in the OAI registry




(a) *All* (b) *Except 'oai.repec'*

FIGURE 4.2: Number of publications by data provider

[52] by requesting *ListRecords* with an updated *resumption token*. Then, it creates a *csv* file for each repository recording how many times any of DCMI terms previously mentioned, such as *identifier*, *publisher*, *subject*, *creator*, *title*, etc, is encoded as a URI, as a TEXT or as a DATE for each publication. The followed criteria to classify the value of a term is really simple: a term will be classified as a DATE when it is presented according to ISO-8601, as a URI when it complies with the RFC-2396 normative, and as a TEXT for the rest of cases.

After that, we have analyzed these files obtaining the following conclusions:

1. ***Down Servers***: Among the 2.656 repositories registered on the *openarchives.org* website, only 1.710 were available and among them, only 1.085 were operative. So, only 40% of registered repositories are currently valid to be harvested.

2. ***Unbalanced Size***: There are providers with less than 50 publications such as the *Annals of National Academy of Medical Sciences Repository (India)* (43 publications) or the *Pediatric Neurology Briefs Repository* (21 publications). Others with more than 150.000 publications such as the *National Library of Australia Digital Object Repository* (178.445 publications) or even more than 300.000 such as the *OAI-PMH gateway for Research Papers in Economics* (oai.repec) (365.419 publications)(*Figure* 4.2).

3. ***Proprietary Date Format***: Trying to discover what is the oldest and the newest publication for each provider we discovered that the format used to present dates is different between providers and even from the recommendations of DCMI [23]. According to DCMI, the recommended best practice is to use the W3CDTF profile of ISO 8601, unfortunately this recommendation sometimes is not followed.

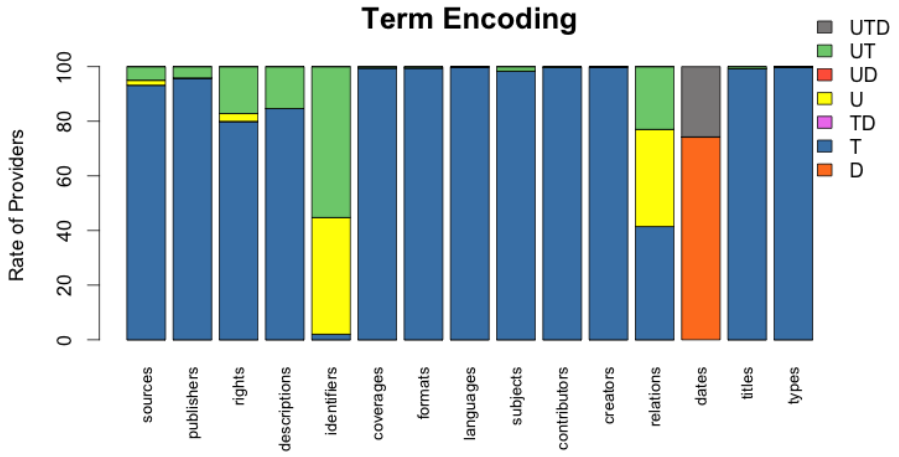



FIGURE 4.3: Distribution of data providers by the term encoding scheme used

4. ***Terms Encoding***: For each term we have calculated all permutations according to the encode scheme used, i.e DATE, URI and/or TEXT. Table 4.1 and figure 4.3 show rates of providers using one of the following schemas: *D* when it is only encoded as DATE; *T* when its is only encoded as TEXT; *U* when it is only encoded as URI; *TD* when it is encoded both as TEXT and/or as DATE; *UT* when it is encoded both as TEXT and/or URI; *UD* when it is encoded both as DATE and/or URI; *UTD* when it is encoded both as URI and/or DATE and/or TEXT.

   As can be seen, *diversity* is the word that best represents how terms are used in the OAI-PMH data providers. As could be expected, *date* is usually encoded as a DATE but sometimes (25.73%) also appear as a URI or as a TEXT. In our opinion, this is not strictly truth. A more detailed analysis showed that this ratio reflects records including the label `dc:date` empty.

   The most frequent schema used by providers to encode the DCMI terms is TEXT, except to represent an *identifier*, which is usually encoded as URI, and *date*, as before mentioned encoded as DATE. It might be thought that this behaviour is the same to express *relations*, however it is also encoded (many times) as TEXT.

   It should also be highlighted that *source*, *rights*, *publishers* and *subjects*, although rarely, appear encoded as URI. We think that this behaviour should be much more extensive to handle correctly references to these entities. However, *description*, which is defined as an abstract, a table of contents, a graphical representation, or a free-text account of the resource, also appears encoded as URI and, in our opinion, it should be only encoded as TEXT.

| **Term** | **D** | **T** | **T+D** | **U** | **U+D** | **U+T** | **U+T+D** |
|---|---|---|---|---|---|---|---|
| *source* | 0% | 93.19% | 0% | 1.83% | 0% | 4.96% | 0% |
| *publisher* | 0% | 95.77% | 0.09% | 0% | 0% | 4.13% | 0% |
| *rights* | 0% | 79.87% | 0.09% | 2.84% | 0% | 17.18% | 0% |
| *description* | 0% | 84.65% | 0% | 0% | 0% | 15.34% | 0% |
| *identifier* | 0% | 2.02% | 0% | 42.73% | 0% | 55.23% | 0% |
| *coverage* | 0% | 99.54% | 0% | 0% | 0% | 0.45% | 0% |
| *format* | 0% | 99.54% | 0% | 0% | 0% | 0.45% | 0% |
| *language* | 0% | 99.90% | 0% | 0% | 0% | 0.09% | 0% |
| *subject* | 0% | 98.34% | 0% | 0% | 0% | 1.65% | 0% |
| *contributor* | 0% | 99.72% | 0% | 0% | 0% | 0.27% | 0% |
| *creator* | 0% | 99.90% | 0% | 0% | 0% | 0.09% | 0% |
| *relation* | 0% | 41.45% | 0% | 35.56% | 0% | 22.97% | 0% |
| *date* | 74.17% | 0% | 0.09% | 0% | 0% | 0% | 25.73% |
| *title* | 0% | 99.26% | 0% | 0% | 0% | 0.73% | 0% |
| *type* | 0% | 99.81% | 0% | 0.09% | 0% | 0.09% | 0% |

TABLE 4.1: Rate of providers by encoding schema (Date, Text, Uri) used for DCMI terms

5. ***Terms Frequency***: Also, we have counted how many times a term appears in the same record of a publication. Grouping this information by data provider and calculating the rate of them that present one of these behaviours: *never*, *one* or *more* than one, we have built table 4.2.

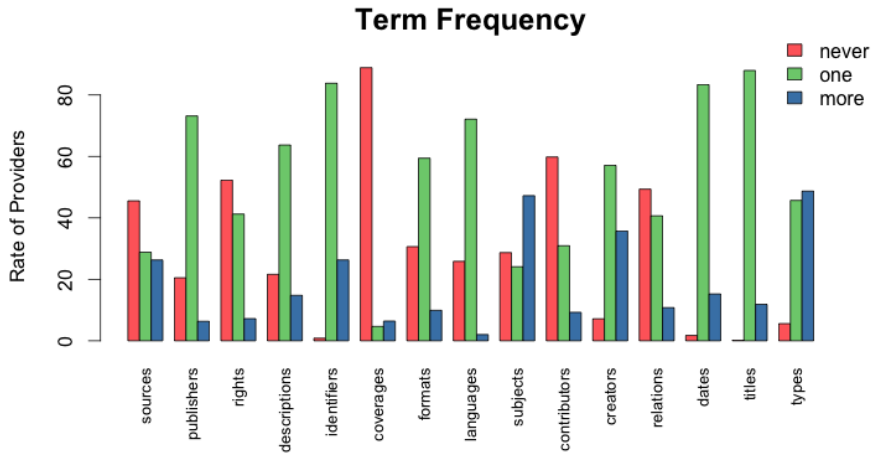



FIGURE 4.4: Distribution of data providers by the term frequency used

As shown in that table, and in *Figure* 4.4, the least frequently used terms are *coverage*, *contributor*, *relation* and *rights*. This may show two points of view: research resources are usually isolated (neither *contributors* from other areas nor *relation*

to other research resources), or research resources published without enough information such as *rights* (privileges) or *coverage* (spatial or temporal topic). In contrast, the most frequently used terms are *title*, *identifier* and *date*. This was to be expected because it is the minimum information to describe a resource.

It is interesting that the *publisher* term was not the highest value. Some years ago a research need a *publisher* to be published, but nowadays it could be distributed independently. This pattern is reflected in the rate of providers that always include that term.

Other interesting tendency is showed by the high value of providers that include more than one *type*. Although this value is not always included, it usually appears more than one time reflecting the multi-format of the resource.

Curiously, there are providers that, at least once, have published resources with more than one *title*, or more than one *description*, or even more than one *subject*. The reason is because the base language is different to english and it appears translated to english as the *Listing* 4.1 shows:

```
<metadata>
  <dc>
   <dc:title xml:lang="de-DE">Urbanitat nach exklusivem Rezept. Die Ausdeutung des
     Stadtischen durch hochpreisige Immobilienprojekte in Berlin und Los Angeles</
     dc:title>
   <dc:title xml:lang="en-US">Urbanity According to an Exclusive Recipe: The
     Interpretation of the Urban through Upscale Real Estate Projects in Berlin and Los
     Angeles</dc:title>
   <dc:subject xml:lang="de-DE">Geographie; Politische Okonomie; Stadtplanung;
     Soziologie</dc:subject>
   <dc:subject xml:lang="de-DE">Gentrifizierung; Urbanitat; Diskurs; Lebensstil</
     dc:subject>
    <dc:subject xml:lang="en-US"></dc:subject>
    <dc:subject xml:lang="en-US">Gentrification; Urbanity; Discourse; Lifestyle</
     dc:subject>
    ...
  </dc>
</metadata>
```

LISTING 4.1: Record with more than one title

6. ***Terms Ambiguity***: The meaning of a term given by a provider may be different from another. For example, some providers consider the url to download the resource (usually *pdf* file) encoded in the term *relation*, e.g *oa.upm.es*. However, others include that information in the term *identifier*, e.g. *eprints.ucm.es*, and others do not even include directly that url, but they include the url of a web page where you can read and download that resource, e.g *sciencepubco*. That is just an example of how this heterogeneity can make developing an unified client difficult.

| Term | Never | One | More |
|---|---|---|---|
| *source* | 45.54% | 28.81% | 26.29% |
| *publisher* | 20.50% | 73.17% | 6.33% |
| *rights* | 52.27% | 41.24% | 7.20% |
| *description* | 21.61% | 63.71% | 14.78% |
| *identifier* | 0.84% | 83.80% | 26.30% |
| *coverage* | 88.88% | 4.66% | 6.41% |
| *format* | 30.63% | 59.42% | 9.90% |
| *language* | 25.85% | 72.11% | 2.02% |
| *subject* | 28.69% | 24.09% | 47.23% |
| *contributor* | 59.80% | 30.92% | 9.25% |
| *creator* | 7.14% | 57.10% | 35.72% |
| *relation* | 49.34% | 40.66% | 10.78% |
| *date* | 1.74% | 83.36% | 15.24% |
| *title* | 0.14% | 87.97% | 11.84% |
| *type* | 5.58% | 45.68% | 48.70% |

TABLE 4.2: Rate of providers by times that DCMI terms appear

## 4.2 Hoarder

Taking into account these considerations, we designed a flexible enough OAI-PMH/RSS Hoarder client [11] to correctly handle these particularities. In the scope of this thesis, we have considered as source of information both OAI-PMH Data Providers, e.g. *openarchives.org* and RSS Sites, e.g. *rss.slashdot.org*. The inherent idea is to force the application to handle different protocol specifications based on common steps typical for data providers. Both RSS and OAI-PMH providers should publish the meta-information using dublin-core terms and also publish it in a passive way, i.e. waiting for a previous request defining a temporal window to list all the available resources. However, OAI-PMH Data Providers usually publish related files (pdf, doc, etc) that have to be downloaded separately from the meta-information, while RSS Sites always include the content and the meta-information in the same resource.

Taking into account these features, the first step was to choose an integration framework that enabled us to represent these differences and to reuse the common actions. In Java there are mainly three frameworks: Spring-Integration [67], Mule-Enterprise Service Bus (Mule-ESB) [51] and Apache Camel [4] with these features. All these solutions are known to implement the well-known *Enterprise Integration Patterns* (EIP) [32] offering a standardized domain-specific language to integrate applications and a set of adapters

to support integration with common external services such as FTP, WS-REST, AMQP services and so on.

Taking into account that they all are open-source and they have special features such as error handling, transactions, multi-threading, scalability and monitoring, and so on, the really discriminative aspects between them are the number of supported technologies, the *domain-specified language* (DSL) used and the facilities to create new adapters. In fact, our *hoarder* has to communicate with OAI-PMH providers, so we need to develop a new OAI-PMH adapter. For this reason, the procedure to create new adapters is a key feature for us, as well as having a large number of existing components easy-to-integrate with other frameworks. Finally, we decided to use Apache Camel [4] as our integration framework to defines routes and workflows.

### 4.2.1 OAI-PMH Camel Component

A new camel component for polling OAI-PMH Data Providers [10] was developed to implement the OAI-PMH Protocol described in Section 2. It allows to create Camel routes with an OAI-PMH Provider as source and also converts from *OAIPMHType* to XML format (*marshal*) or from XML to *OAIPMHType* format (*unmarshal*) using the *OAI-PMH XML Schema definition* [OpenArchive]. The purpose of this feature is to make it possible to use Camel's built-in expressions for manipulating OAI-PMH messages. As shown below, an XPath (*XML Path Language*) [70] expression can also be used to filter the OAI-PMH message::

```
1 from("oaipmh://oa.upm.es/perl/oai2?delay=60000").unmarshal().jaxb("es.upm.oeg.camel.oaipmh.
      model").filter().xpath("//item/request/set[contains(.,'physics')]").to("mock:result");
```

Listing 4.2: Camel route for polling the *UPM* OAI-PMH Data Provider

We have defined some options to customize the behaviour of this component:

- `delay`: Delay in milliseconds between each poll.
- `initialDelay`: Milliseconds before polling starts.
- `userFixedDelay`: Set to true to use fixed delay between pools, otherwise fixed rate is used.
- `verb`: OAI-PMH command such as *ListRecords*, *ListIdentifiers*, *Identify*..

- `metadataPrefix`: Specifies the metadataPrefix of the format that should be included in the metadata part of the returned records.
- `from`: Specifies a lower bound for datestamp-based selective harvesting. UTC DateTime value. After first request, this value is updated to current time if no upper bound is defined
- `until`: Specifies an upper bound for datestamp-based selective harvesting. UTC DateTime value.

### 4.2.2 Camel Routes

Once the OAI-PMH Camel component has been developed, it can be used from a camel route to request the list of records to a Data Provider. After that, we need to understand the obtained record to handle the values of the terms and download the related resource/s. As mentioned before, it has to handle term ambiguity and multi-encoding into the same protocol (RSS or OAI-PMH), so it should be adaptable enough for each DC term used. Usually, the exchanged information is encoded as XML, so we also incorporated *XML Path Language* (XPATH) [70] here to adapt the expressions to read the term values in our routes to handle the heterogeneity showed before. In future cases where *JSON* will be the encoding used by servers, we will use *JSONPATH* instead of *XPATH* to define these expressions.

A new route should be created for each new data provider handled. So, we designed a wrapper for Camel context using the Groovy language [21], to facilitate the way to create/modify OAI-PMH and RSS routes. It allows route definition using a Java style, i.e. using the Camel Domain Specific Language (DSL) directly for creating EIP or routes in a fluent builder style. To reduce the url definitions, we included the following namespaces in the root context:

- **oai**: "http://www.openarchives.org/OAI/2.0/"
- **dc**: "http://purl.org/dc/elements/1.1/"
- **provenance**: "http://www.openarchives.org/OAI/2.0/provenance"
- **oai_dc**: "http://www.openarchives.org/OAI/2.0/oai_dc"

- **rss**: "http://purl.org/rss/1.0/"

Thus, a route for polling an OAI-PMH data provider, create a metadata file for each record and download the related files is:

```
from("oaipmh://oa.upm.es/perl/oai2?initialDelay=1000&delay=60000").
    setProperty(SOURCE_NAME, constant("upm")).
    setProperty(SOURCE_URL, constant("http://oa.upm.es/perl/oai2")).
    to("direct:setCommonOaipmhXpathExpressions").
    setProperty(PUBLICATION_URL, xpath("//oai:metadata/oai:dc/dc:relation/text()",String.class)
      .namespaces(ns)).
    to("direct:retrieveByHttpAndSave")
```

LISTING 4.3: OAI-PMH hoarder route for the *UPM* website

The same for a RSS site:

```
from("rss:http://rss.slashdot.org/Slashdot/slashdot?" +
    "splitEntries=true&consumer.initialDelay=1000&consumer.delay=2000" +
    "&feedHeader=false&filter=true").marshal().rss().
    setProperty(SOURCE_NAME, constant("slashdot")).
    setProperty(SOURCE_URL,  constant("http://rss.slashdot.org/Slashdot/slashdot")).
    to("direct:setCommonRssXpathExpressions").
    to("direct:retrieveByHttpAndSave")
```

LISTING 4.4: RSS hoarder route for the *Slashdot* website

In these cases, some common routes, i.e. set of actions, have been used to encapsulate groups of operations such as `direct:setCommonOaipmhXpathExpressions`, `direct:setCommonRssXpathExpressions` and `direct:retrieveByHttpAndSave`.

A camel route defines flows of works starting in a `from` point and deriving to one or more output flows (`to`). These common flows define common actions grouped in a same starting point (`from`).

In such a way, `direct:setCommonOaipmhXpathExpressions` defines the set of expressions, usually *XPATH* expressions, to extract the values of the Dublin-Core terms from an OAI-PMH record as follows:

```
from("direct:setCommonOaipmhXpathExpressions").
    setProperty(SOURCE_PROTOCOL,
     constant("oaipmh")).
    setProperty(SOURCE_URI,
     simple("http://www.epnoi.org/oaipmh/${property." + SOURCE_NAME + "}")).
    setProperty(PUBLICATION_TITLE,
     xpath("//oai:metadata/oai:dc/dc:title/text()",String.class).namespaces(ns)).
    setProperty(PUBLICATION_DESCRIPTION,
     xpath("//oai:metadata/oai:dc/dc:description/text()",String.class).namespaces(ns)).
    setProperty(PUBLICATION_PUBLISHED,
     xpath("//oai:header/oai:datestamp/text()",String.class).namespaces(ns)).
    setProperty(PUBLICATION_URI,
     xpath("//oai:header/oai:identifier/text()",String.class).namespaces(ns)).
    setProperty(PUBLICATION_URL,
```

```
15      xpath("//oai:metadata/oai:dc/dc:identifier/text()",String.class).namespaces(ns)).
16    setProperty(PUBLICATION_LANGUAGE,
17      xpath("//oai:metadata/oai:dc/dc:language/text()",String.class).namespaces(ns)).
18    setProperty(PUBLICATION_RIGHTS,
19      xpath("//oai:metadata/oai:dc/dc:rights/text()",String.class).namespaces(ns)).
20    setProperty(PUBLICATION_CREATORS,
21      xpath("string-join(//oai:metadata/oai:dc/dc:creator/text(),\";\")",String.class).
        namespaces(ns)).
22    setProperty(PUBLICATION_FORMAT,
23      xpath("substring-after(//oai:metadata/oai:dc/dc:format/text(),\"/\")", String.class).
        namespaces(ns)).
24    setProperty(PUBLICATION_METADATA_FORMAT,
25      constant("xml")).
26    to(direct:avoidDeletedMessages);
```

LISTING 4.5: variable expressions to understand OAI-PMH terms

Some of them can be modified for a specific data provider, for instance for *oa.upm.es* data provider the `PUBLICATION_URL` variable requires a specific expression so it is overriden later as showed in Listing 4.3.

In a similar way, the RSS data provider requires a common `direct:setCommonRssXpathExpressions` as follows:

```
1  from("direct:setCommonRssXpathExpressions").
2     setProperty(SOURCE_PROTOCOL,
3       constant("rss")).
4     setProperty(SOURCE_URI,
5       simple("http://www.epnoi.org/rss/${property."+SOURCE_NAME+"}")).
6     setProperty(PUBLICATION_TITLE,
7       xpath("//rss:item/rss:title/text()", String.class).namespaces(ns)).
8     setProperty(PUBLICATION_DESCRIPTION,
9       xpath("//rss:item/rss:description/text()", String.class).namespaces(ns)).
10    setProperty(PUBLICATION_PUBLISHED,
11      xpath("//rss:item/dc:date/text()", String.class).namespaces(ns)).
12    setProperty(PUBLICATION_URI,
13      xpath("//rss:item/rss:link/text()", String.class).namespaces(ns)).
14    setProperty(PUBLICATION_URL,
15      xpath("//rss:item/rss:link/text()", String.class).namespaces(ns)).
16    setProperty(PUBLICATION_LANGUAGE,
17      xpath("//rss:channel/dc:language/text()", String.class).namespaces(ns)).
18    setProperty(PUBLICATION_RIGHTS,
19      xpath("//rss:channel/dc:rights/text()", String.class).namespaces(ns)).
20    setProperty(PUBLICATION_CREATORS,
21      xpath("string-join(//rss:channel/dc:creator/text(),\";\")", String.class).namespaces(ns)).
22    setProperty(PUBLICATION_FORMAT,
23      constant("htm")).
24    setProperty(PUBLICATION_METADATA_FORMAT,
25      constant("xml"));
```

LISTING 4.6: variable expressions to understand RSS terms

In addition, other common flows have been developed such as *direct:avoidDeleted*, *direct:saveToFile*, *direct:downloadByHttp* and *direct:downloadByHttpAndSave* to be reused from high-level routes such as *direct:setCommonRssXpathExpressions* or *direct:setCommonOaipmhXpathExpressions*.

#### 4.2.2.1 direct:avoidDeleted

This route avoids to process an OAI-PMH record when it has been removed. In these cases, according to the protocol definition, the XML includes the attribute `status` equals to `deleted` in the header of the record:

```
from("direct:avoidDeletedMessages").
  choice().
  when().xpath("//oai:header[@status=\"deleted\"]", String.class, ns).stop().
  end();
```

LISTING 4.7: camel route to avoid processing OAI-PMH deleted records

#### 4.2.2.2 direct:saveToFile

This route creates a file, reusing the existing *file* Camel component, in a path based on some properties such as SOURCE_PROTOCOL and SOURCE_NAME with the content received in the route. Camel defines a *Body* and a *Header* for all its internal messages, so in this case, the *Body* of the message (serialized as XML) will be written in a file. This route is used to create both the metadata and the related files:

```
from("direct:saveToFile").
  setHeader(ARGUMENT_PATH, simple("${property." + SOURCE_PROTOCOL + "}/${property." +
    SOURCE_NAME + "}/${property" + PUBLICATION_PUBLISHED_DATE + "}/${header." + ARGUMENT_NAME
    + "}")).
  log(LoggingLevel.INFO,LOG,"File Saved: '${header."+ARGUMENT_PATH+"}'").
  to("file:"+basedir+"/?fileName=${header."+ARGUMENT_PATH+"}");
```

LISTING 4.8: camel route to save a file

#### 4.2.2.3 direct:downloadByHttp

This route downloads a web resource from the URL defined in the property *ARGUMENT_PATH*. This external resource will be serialized as XML in the *Body* of the internal Camel message. As showed above, it uses the existing *http* camel component to connect and download the remote resource. This route is used to obtain the related files of an OAI-PMH record:

```
from("direct:downloadByHttp").
   // Filter resources with available url
   filter(header(ARGUMENT_PATH).isNotEqualTo("")).
   setHeader(Exchange.HTTP_METHOD, constant("GET")).
   setHeader(Exchange.HTTP_URI, simple("${header." + ARGUMENT_PATH + "}")).
   to("http://dummyhost?throwExceptionOnFailure=false");
```

LISTING 4.9: camel route to download a web resource

#### 4.2.2.4 direct:retrieveByHttpAndSave

This route combines the previous routes to save both the metadata and the related files from an OAI-PMH record or a RSS new:

```
from("direct:retrieveByHttpAndSave").
  process(timeClock).
  process(uuidGenerator).
  setHeader(ARGUMENT_NAME,
    simple("${property."+PUBLICATION_UUID+"}."+"${property."+PUBLICATION_METADATA_FORMAT+"}"))
    .
  to("direct:saveToFile").
  setHeader(ARGUMENT_PATH,
    simple("${property."+PUBLICATION_URL+"}")).
  to("direct:downloadByHttp").
  setHeader(ARGUMENT_NAME,
    simple("${property."+PUBLICATION_UUID+"}."+"${property."+PUBLICATION_FORMAT+"}")).
  to("direct:saveToFile").
  setProperty(PUBLICATION_URL_LOCAL,
    simple("${header." + ARGUMENT_PATH + "}"));
```

LISTING 4.10: camel route to save both the metadata and the related files

### 4.2.3 Evaluation

In this section, a real use case is presented to retrieve resources from a set of OAI-PMH data providers listed in the OpenArchive website [52]. Our *Hoarder* [11] is a Java application wrapped with *Java Service Wrapper* (JSW) to easily be executed as a service in any operating system. For this evaluation, a MacOSX environment (Yosemite 10.10.3) was setup in an Intel Core i7 2,3Ghz (x8) with 16 GB of memory.

The following three steps detail how to create a new route in the application to check and extract content from a data provider:

1. First, obtain the base URL and the ID of the repository from the registration list supported by OpenArchives [52].

2. After that, perform a *listOfRecords* request to the OAI-PMH service of the repository to check how and where the url to download the related resource appears in the records. It may appear in the *relation* section or in the *identifier* section, or even it may appear referencing an online viewer instead of a downloadable resource, requiring to be modified to download the resource. In future versions, this manual step will be omitted because it will be automatically performed.

3. Finally, create a workflow expressed as a Camel [4] route including the base URL, the OAI ID and the XPATH expressions to adapt the extraction of values to the particularities of the provider.

Thus, trying to compose a varied research corpus, we decided to include some OAI-PMH data providers published by Innovare Journal [40] oriented to different research areas such as Agricultural Science (IJAGS), Business Management (IJBM), Education (IJOE), Ayurvedic Science (IJAS), Engineering and Technology (IJET), Health Science (IJHS), Life Science (IJLS), Medical Science (IJMS), Social Science (IJSS) and Science (IJS). In all these services, the url to download the related resource (usually a *pdf* file) references to an online viewer instead of the file, so the route added to our *hoarder* application should include an expression that composes a valid url pointing directly to the resource from the url pointing to the viewer.

For example, once we have obtained the url and the id of the Innovare Journal of Agricultural Science repository from the OpenArchives website, we check the records returned by the OAI service after request a *listOfRecords*. The following piece of code show one of them:

http://innovareacademics.in/journals/index.php/ijags/oai?verb=ListRecords&metadataPrefix=oai_dc:

```
<record>
    <header>
        <identifier>oai:ojs.innovareacademics.in:article/4366</identifier>
        <datestamp>2015-05-14T14:56:00Z</datestamp>
        <setSpec>ijags:Res</setSpec>
    </header>
    <metadata>
        <oai_dc:dc
            xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
            xmlns:dc="http://purl.org/dc/elements/1.1/"
            xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
  http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
            <dc:title xml:lang="en-US">Optimization of Irrigation and Fertilizer for Sweet
    Corn (Zea mays L. var. saccharata Sturt) under Climate Change Conditions</dc:title>
            <dc:creator>Mathukia, R. K.; Associate Research Scientist, Department of Agronomy,
     College of Agriculture, Junagadh Agricultural University, Junagadh-362001 (Gujarat)</
    dc:creator>
            <dc:description xml:lang="en-US">A field experiment was conducted ...</
    dc:description>
            <dc:publisher xml:lang="en-US">Innovare Journal of Agricultural Sciences</
    dc:publisher>
            <dc:contributor xml:lang="en-US">Junagadh Agricultural University</dc:contributor>
            <dc:date>2015-01-14</dc:date>
            <dc:type>info:eu-repo/semantics/article</dc:type>
            <dc:type>info:eu-repo/semantics/publishedVersion</dc:type>
            <dc:type xml:lang="en-US"></dc:type>
```

```
23              <dc:format>application/pdf</dc:format>
24              <dc:identifier>http://innovareacademics.in/journals/index.php/ijags/article/view
      /4366</dc:identifier>
25              <dc:source xml:lang="en-US">Innovare Journal of Agricultural Sciences; Vol 3 Issue
       1 2015 ( January - March )</dc:source>
26              <dc:source>2321-6832</dc:source>
27              <dc:language>eng</dc:language>
28              <dc:relation>http://innovareacademics.in/journals/index.php/ijags/article/view
      /4366/1876</dc:relation>
29          </oai_dc:dc>
30      </metadata>
31  </record>
```

Listing 4.11: OAI-PMH record from the IJAGS repository

The url to download the related resource appears in the **dc:identifier** section and it is pointing to an online viewer instead of the file. So we need to create a XPATH expression to compose the valid url. In this case it is not really complex because replacing *view* by *download* in the url is enough to download the resource, so the workflow created in our Hoarder is as follows::

```
1  from("oaipmh://innovareacademics.in/journals/index.php/ijags/oai?
2    initialDelay=1000&delay=60000").
3      setProperty(SOURCE_NAME, constant("ijags")).
4      setProperty(SOURCE_URL, constant("http://innovareacademics.in/journals/index.php/ijags/oai
       ")).
5      to("direct:setCommonOaipmhXpathExpressions").
6      setProperty(PUBLICATION_URL,
7          xpath("replace(substring-before(concat(string-join(//oai:metadata/oai:dc/dc:relation/
       text(),\";\"),\";\"),\";\"),\"view\",\"download\")",String.class).namespaces(ns)).
8      to("direct:retrieveByHttpAndSave")
```

Listing 4.12: camel route to extract resources from the IJAGS repository

As showed in the code, some XPATH operations such as *substring-before*, *concat* and *string-join* have been used to compose an url that allow hoarder to download the related resource from the OAI-PMH metadata. Depending on the type of repository, this expression can be more or less complex.

More than 200 resources were downloaded from these 10 data providers, 100 of them were selected to create our research corpus taking 10 resources from each provider. The list of these resources is detailed in Appendix D.

## 4.3 Harvester

As showed in figure 4.1, the *harvester* application [8] creates research objects from both metadata and related files previously downloaded by the *hoarder* client. This application

was developed using Camel to define the harvesting flows.

Now, the goal is to build research objects containing both meta-information and list of words extracted from the content of the resource or resources. The Apache PDF parser (*PDFTextStripper*) was used to extract the text from the *pdf* files along with the *Lucene* classifier to obtain the *bag-of-words* from a resource file. After that, a *JSON* parser (*GSON*) was required to create the related *research object* in a predefined *JSON* scheme. Actually, it includes the following information:

- **uri**: global identifier of the resource.
- **url**: local path to the file containing the resource or the resources. It may be a *pdf* file, i.e. research object composed by only one resource, or a *zip* file, i.e. research object containing several resources.
- **source**: some details about the data provider where the resources were downloaded such as name, uri, url and protocol, e.g *oaipmh* or *rss*.
- **meta-information**: some extra information about the resource such as title, publish date, format, language, rights and description. It also contains the list of authors, described by the name, the surname and a URI.
- **bagOfWords**: list of word stems generated from the content of related resources, e.g. pdf files. For non textual resources, the annotated information is used.
- **resources**: list of other nested research objects.

The *Harvester* application consumes files from a folder where the *Hoarder* application drops them directly, in real time. It is possible using the *doneFileName* option included in Camel that extends the limited functionality defined in the File IO Api of JDK. Thus, in a continuous flow of work, the *Hoarder* downloads resources and meta-information from data providers to a shared folder where the *Harvester* takes them and generates research objects from them to other shared folder to be processed by RESSIST (Figure 4.5).

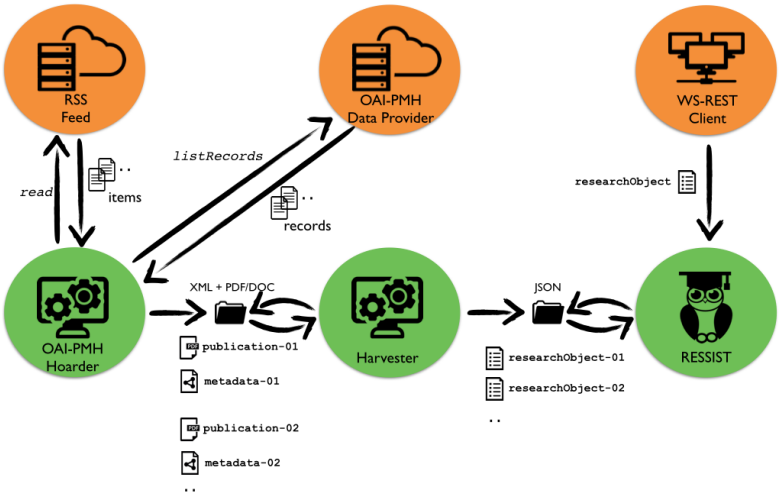



FIGURE 4.5: Workflow architecture

### 4.3.1 Evaluation

Using the resources previously downloaded by the *Hoarder* application, the *Harvester* application creates research objects based on them.

For example, the *Hoarder* application has downloaded the following meta-information (XML file) along with the publication (pdf file):

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:provenance="http://www.
    openarchives.org/OAI/2.0/provenance" xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/
    oai_dc/" xmlns:dc="http://purl.org/dc/elements/1.1/">
    <responseDate>2015-06-23T08:34:48Z</responseDate>
    <request verb="ListRecords" metadataPrefix="oai_dc">http://innovareacademics.in/journals/
    index.php/ijss/oai</request>
    <ListRecords>
        <record>
            <header>
                <identifier>oai:ojs.innovareacademics.in:article/1185</identifier>
                <datestamp>2014-07-11T16:00:21Z</datestamp>
                <setSpec>ijss:ART</setSpec>
            </header>
            <metadata>
                <dc>
                    <dc:title xml:lang="en-US">PRELIMINARY STUDY ON THE USE OF SOUND AND
    ACOUSTICS IN IGBO CULTURAL COMMUNICATION</dc:title>
                    <dc:creator>Ahamefula, Ndubuisi Ogbonna; Department of Linguistics, Igbo
    and Other Nigerian Languages, University of Nigeria, Nsukka
Acoustical Society of Nigeria
Linguistic Association of Nigeria
Igbo Studies Association
West African Linguistic Society
The Linguist List</dc:creator>
                    <dc:creator>Okoye, Chinenye L.</dc:creator>
                    <dc:creator>Onwuegbuchunam, Marcellus O.</dc:creator>
                    <dc:creator>Uzoigwe, Benita C.</dc:creator>
                    <dc:creator>Nneji, Ogechukwu M.</dc:creator>
                    <dc:description xml:lang="en-US">Language had been a veritable tool for
    communication among ...</dc:description>
```

```
                    <dc:publisher xml:lang="en-US">Innovare Journal of Social Sciences</dc:publisher>
                    <dc:contributor xml:lang="en-US"></dc:contributor>
                    <dc:date>2014-07-01</dc:date>
                    <dc:type>info:eu-repo/semantics/article</dc:type>
                    <dc:type>info:eu-repo/semantics/publishedVersion</dc:type>
                    <dc:type xml:lang="en-US">Peer-reviewed Article</dc:type>
                    <dc:format>application/pdf</dc:format>
                    <dc:identifier>http://innovareacademics.in/journals/index.php/ijss/article/view/1185</dc:identifier>
                    <dc:source xml:lang="en-US">Innovare Journal of Social Sciences; Vol 2 Issue 3 2014 (July-September)</dc:source>
                    <dc:language>eng</dc:language>
                    <dc:relation>http://innovareacademics.in/journals/index.php/ijss/article/view/1185/786</dc:relation>
                </dc>
            </metadata>
        </record>
    </ListRecords>
</OAI-PMH>
```

LISTING 4.13: Metadata file

Then, the *Harvester* application analyzes both that metadata file (XML) and its content file (PDF) to compose a *json* file describing the research object:

```
{
    "uri": "oai:ojs.innovareacademics.in:article/1185",
    "url": "file:///opt/drinventor/epnoi-hoarder-1.0.9/publications/oaipmh/ijss/2014-07-11/f863f27c-2451-409f-8771-c12433d1a4d9.pdf",
    "source": {
        "name": "innovareacademics.in",
        "uri": "http://www.epnoi.org/oaipmh/innovareacademics.in/journals/index.php/ijss/oai",
        "url": "http://innovareacademics.in/journals/index.php/ijss/oai",
        "protocol": "oaipmh"
    },
    "metainformation": {
        "title": "PRELIMINARY STUDY ON THE USE OF SOUND AND ACOUSTICS IN IGBO CULTURAL COMMUNICATION",
        "published": "2014-07-11T16:00:21Z",
        "format": "pdf",
        "language": "eng",
        "rights": "",
        "description": "Language had been a veritable tool for communication among ...",
        "creators": [
            {
                "uri": "http://resources.ressist.es/author/Ahamefula-Ndubuisi\%20Ogbonna",
                "name": "Ndubuisi Ogbonna",
                "surname": "Ahamefula"
            },
            {
                "uri": "http://resources.ressist.es/author/Department\%20of\%20Linguistics-Igbo\%20and\%20Other\%20Nigerian\%20Languages",
                "name": "Igbo and Other Nigerian Languages",
                "surname": "Department of Linguistics"
            },
            {
                "uri": "http://resources.ressist.es/author/Okoye-Chinenye\%20\%20L.",
                "name": "Chinenye L.",
                "surname": "Okoye"
            },
            {
                "uri": "http://resources.ressist.es/author/Onwuegbuchunam-Marcellus\%20O.",
```

```
35                  "name": "Marcellus O.",
36                  "surname": "Onwuegbuchunam"
37              },
38              {
39                  "uri": "http://resources.ressist.es/author/Uzoigwe-Benita\%20C.",
40                  "name": "Benita C.",
41                  "surname": "Uzoigwe"
42              },
43              {
44                  "uri": "http://resources.ressist.es/author/Nneji-Ogechukwu\%20M.",
45                  "name": "Ogechukwu M.",
46                  "surname": "Nneji"
47              }
48          ]
49      },
50      "bagOfWords": [
51          "commun",
52          "sound",
53          "cultur",
54          "instrument",
55          "music",
56          "other",
57          "announc",
58          "languag",
59          ...
60      ]
61  }
```

Listing 4.14: Research Object as *JSON* string

As showed before, some unexpected values appear in the research object such as the *author "Igbo and Other Nigerian Languages"* which is really a description of the previous author *"Ndubuisi Ogbonna"* that was included in the same **dc:creator** section. Maybe defining a guide of good practices to create meta files using Dublin-Core annotation will be helpful for this type of posterior analysis.

Finally, the *license rights* about the publication are not included and the *bagOfWords* was created using a stemming process based on Lucene indexing of the *pdf* file. Future works will later execute the stemming process over the list of concepts created from the list of words extracted from the publication file or files.

# Chapter 5

# Contribution 2: LDA configuration using evolutionary multi-objective optimization

As mentioned in Section 2, an LDA model is used to describe the inherent topic distribution of existing *research objects*. This model requires some parameters to be built, and they need to be adjusted to obtain a high quality model.

Since LDA is characterized by Dirichlet distributions of topics and documents, i.e. multivariate generalization of the Beta distribution, it is parameterized by two positive shape parameters, $\alpha$ and $\beta$, that appear as exponents of the random variable and control the shape of the distribution. Moreover, the dimensionality of each Dirichlet distribution has to be fixed. So the dimensionality value of the Dirichlet distribution of topics is known and equals to the size of the vocabulary. However, the dimensionality of the Dirichlet distribution of documents, i.e. number of topics, is assumed to be known and fixed.

Thus, we need to estimate three parameters: the *number of topics (k)*, the concentration parameter ($\alpha$) for the prior placed on documents' distributions over topics and the concentration parameter ($\beta$) for the prior placed on topics' distributions over terms . Some authors [7] have proposed inferences to calculate these parameters, however the implementation of LDA made by Spark (based on *Expectation/Maximization*) and used by RESSIST does not admit these values yet.

In addition, all parameters are corpus-level parameters, so we need to calculate new values whenever the corpus changes. From the point of view of efficiency, this operation is executed in background mode each time a group of resources are added. The size of that group is defined beforehand.

## 5.1 Multi-Objective Evolutionary Approach

New values of *log-likelihood* and *log-prior* are obtained for each new LDA execution that measure the goodness of the model. The higher these values are, the better the model fits. For this reason, having several conflicting objetives, i.e. improvement of one objective may lead to deterioration of another, and having parameters to estimate ($k$, $\alpha$ and $\beta$), a *Multi-Objetive Evolutionary Algorithm (MOEA)* is used to find the *Pareto* optimal solution.

A single solution, which optimizes *log-likelihood* and *log-prior* simultaneously does not exist. Instead, the best trade-off solution called *Pareto optimal* will be obtained. Taking into account performance behaviour [76] and to prevent new objectives derived from the use of the model, the *Non-Sorting Genetic Algorithm-III (NSGA-III)* [27] is chosen and the optimization problem is defined as follows:

- **Objectives**:
  - $\min \| \log(likelihood) \|$
  - $\min \| \log(prior) \|$
- **Constraints**:
  - $5.1 < \alpha < 20.0$ : Document Concentration. It represents the distribution of a document in topics. That is, how specific is a document. The lower boundary of $\alpha$ (5.1) is greater than the higher boundary of $\beta$ (5.0) because, in our opinion, if a term belongs to more than one topic, a document will contain equal or more number of topics than those contained in the term. Moreover, we have considered than the number of topics in a document is, at least, 4 times greater than the topics contained in a term, for that reason the higher boundary is 20.0 for $\alpha$ and 5.0 for $\beta$.

  - $1.0 < \beta < 5.0$ : Topic Concentration. It represents the distribution of a topic over terms. That is, how a term can belong to several topics. In our opinion, a term can only belong to no more than 5 topics, because greater values will create more ambiguous models.
  - $0 < k < 2 * \sqrt{p/2}$ : Number of topics. Usually around the root square of the half of population (p).

- **Crossover**: Motivated by the success of binary-coded genetic algorithms in problems with discrete search space, the operator selected was *Simulated Binary Crossover (SBX)* [25] that solves problems having a continuous search space instead of binary. This operator has a search power similar to that of the single-point crossover. It was set to 0.9 to facilitate the explorative capacity of the algorithm.

- **Mutation**: Mutation operators have been utilized extensively in MOEAs as solution variation mechanisms. Mutation operators assist to the better exploration of the search space [47]. Different approaches have been proposed depending on the representation used in MOEAs such as binary or real values. In this case, the operator selected was the *Polynomial Mutation* operator [26] [28] which allows big jumps in the search space of the decision variable, escaping from local optima and modifying a solution when on the boundary. It was set to 1.0 to promote the explorative analysis.

- **Selection**: The global best solution is selected by a *N-ary Tournament* operator. This operator prefers feasible solutions over infeasible solutions (for constraint handling), non-dominated solutions over dominated solutions (for handling multiple objetives) and less-crowed solutions over more-crowded solutions (for the maintenance of diversity).

The boundaries of the constraints have been defined, as previously mentioned, according to the implementation of the LDA algorithm made by Spark. The minimum value of the parameters, *alpha* and *beta*, is defined to 1.0 by default, but they will be different in our application. Since the *beta* parameter describes the concentration of topics in words, we consider only low values (lower than 5.0 and greater than 1.0) trying to get more representative words for each topic. Defining this range of values, the algorithm will avoid using the same words to characterize different topics, getting distinguished

distributions of topics in documents. Moreover, defining high values for the *alpha* parameter (between 5.1 and 20.0), the algorithm considers that a document can contain more than one topic, but these distributions will not be smooth. We are looking for a characterization that enables us to handle more than one topic in a document, but also enough differences between the topic distributions of different documents to group documents that are talking about the same area or in the same way.

The number of topics will be between 1 and an empirical value defined by the root square of the half of population ($p$). This approximation is useful to avoid a high exploration during the learning process focusing on a smaller set of values.

## 5.2 Evaluation

Trying to measure the learning process, we executed the evolutionary algorithm implemented by the JMetal framework [41] on our corpus previously created by the *Hoarder* and the *Harvester* applications. We used the implementation of LDA developed by Apache Spark [5] that learns the model using *Expectation-Maximization (EM)* on the likelihood function ($p(O \mid \mu)$). The values of *logLikelihood* and *logPrior* are obtained from that model using the research objects included in the corpus. As aforementioned in Section 4.2.3 and detailed in Appendix D, the corpus is composed by 100 research objects balanced over 10 different research areas: Agricultural Science (IJAGS), Business Management (IJBM), Education (IJOE), Ayurvedic Science (IJAS), Engineering and Technology (IJET), Health Science (IJHS), Life Science (IJLS), Medical Science (IJMS), Social Science (IJSS) and Science (IJS).

A first analysis was to execute the learning process setting a maximum number of executions to 30 and a maximum number of LDA iterations to 20. The rest of values (topics, alpha and beta) were dynamically obtained by the NSGA-III algorithm trying to optimize the final values of *LogLikelihood* and *LogPrior*. The results are showed in table 5.1:

Taking into account the constraints of the parameters, the number of *topics* for this corpus is limited between [1-14] (for a population of 100 individuals, the $2 * \sqrt{p/2}$ is equals to 14) , the *alpha* value between [5.1-20.0] and the *beta* value between [1.1-5.1]. Then, according to the results, the value of *beta* is the most stable, only varying twice

| Test | Topics | Alpha | Beta | LogLikelihood | LogPrior | MaxIters | Time(ms) |
|---|---|---|---|---|---|---|---|
| *1* | 4 | 9.6 | 1.1 | $-392.121,13$ | $-4,40$ | 30/20 | 217.849 |
| *2* | 10 | 7.9 | 1.1 | $-381.553,12$ | $-10,70$ | 30/20 | 277.912 |
| *3* | 6 | 5.1 | 1.1 | $-386.208,24$ | $-6,60$ | 30/20 | 275.010 |
| *4* | 6 | 6.1 | 4.1 | $-418.703,32$ | $-178,14$ | 30/20 | 167.890 |
| *5* | 3 | 6.1 | 1.1 | $-395.317,26$ | $-3,36$ | 30/20 | 310.983 |
| *6* | 11 | 12.1 | 1.1 | $-385.879,43$ | $-11,71$ | 30/20 | 188.025 |
| *7* | 1 | 13.1 | 2.1 | $-411.222,50$ | $-11,32$ | 30/20 | 275.841 |
| *8* | 3 | 7.1 | 1.1 | $-395.034,58$ | $-3,37$ | 30/20 | 364.952 |
| *9* | 8 | 14.4 | 1.1 | $-385.927,65$ | $-8,66$ | 30/20 | 290.385 |
| *10* | 2 | 7.1 | 1.1 | $-400.680,65$ | $-2,24$ | 30/20 | 215.063 |

TABLE 5.1: LDA configurations suggested by the NSGA-III algorithm after 30 evaluations of 20 executions

while the value of *topics* is the most scattered. This behaviour shows that only taking into account the values of *LogLikelihood* and *LogPrior*, a LDA model using 11 topics may have a similar accuracy to another that uses only 6. It occurs because the concentration of topics in a document (*alpha* value) and the concentration of topics in a word (*beta* value) are different in both cases. So, for the LDA model that uses 11 topics, the value of *alpha* is 12.1 and the value of *beta* 1.1, while for the model that uses 6 topics, the value of *alpha* is equal to 6.1 and the value of *beta* is 1.1. Reasoning about this, when the number of topics is low, the concentration of topics in documents is also low, because the topics in that case are more general than when there are more of them. In these cases they are more specific.

As expected, the best configuration defines 10 topics, i.e the same number of different research areas included in the corpus, with a level of concentration of topics in documents (*alpha*) equals to 7.9 and a level of concentration of words in topics (*beta*) equals to 1.1.

All these tests are executed with a maximum number of iterations for NSGA-III equal to 30 and a maximum number of iterations for LDA equal to 20. Trying to discover whether the first value, i.e. max iterations for NSGA-III, can affect to the final result we have executed the same algorithm increasing it to 500. We considered that value because an evolutionary algorithm needs many executions to explore the population. The results are showed in the table 5.2.

Now, the best configuration (highest *loglikelihood*$=-380.974,83$) defines 9 topics, 6.7 of *alpha* concentration and 1.1 of *beta* concentration. However, the parameters show

| Test | Topics | Alpha | Beta | LogLikelihood | LogPrior | MaxIters | Time(ms) |
|---|---|---|---|---|---|---|---|
| *1* | 2 | 6.9 | 1.1 | $-401.676,96$ | $-2,24$ | 500/20 | 671.791 |
| *2* | 2 | 6.1 | 1.3 | $-402.490,28$ | $-6,41$ | 500/20 | 619.366 |
| *3* | 12 | 5.4 | 1.1 | $-378.598,41$ | $-12,7$ | 500/20 | 781.803 |
| *4* | 9 | 6.7 | 1.1 | $-380.974,83$ | $-9,71$ | 500/20 | 870.969 |
| *5* | 7 | 6.2 | 1.1 | $-387.861,83$ | $-7,56$ | 500/20 | 499.501 |

TABLE 5.2: LDA configurations suggested by the NSGA-III algorithm after 500 evaluations of 20 executions

| Test | Topics | Alpha | Beta | LogLikelihood | LogPrior | MaxIters | Time(ms) |
|---|---|---|---|---|---|---|---|
| *1* | 11 | 5.3 | 1.1 | $-377.300,52$ | $-11,74$ | 200/100 | 1.276.498 |
| *2* | 7 | 10.9 | 1.3 | $-394.008,71$ | $-21,36$ | 200/100 | 1.342.360 |
| *3* | 7 | 5.1 | 1.8 | $-401.114,06$ | $-54,60$ | 200/100 | 1.204.178 |
| *4* | 12 | 6.1 | 1.1 | $-376.704,06$ | $-12,72$ | 200/100 | 1.455.811 |
| ***5*** | 12 | 5.9 | 1.1 | $-377.411,04$ | $-12,76$ | 200/100 | 1.167.862 |

TABLE 5.3: LDA configurations suggested by the NSGA-III algorithm after 200 evaluations of 100 executions

the same behaviour than before, being the topics and *alpha* values the most scattered values. No improvement was detected increasing only the maximum number of iterations of NSGA-III, so we decided to increase also the number of iterations of LDA to 100 and reducing the maximum for NSGA-III to 200. Then, we obtained more accurate results, as shown in table 5.3.

These results show that the evolutionary algorithm as well as the LDA Model require a high number of iterations. Now, the values of *LogLikelihood* are usually better than before, and the best configuration appears in the case: 12 topics, *alpha* equals to 5.9 and *beta* equals to 1.1 to create a model with an accurate equals to $-377.411,04$. We used this configuration for the rest of evaluations.

Note the variability of execution time. This is because we introduced a small cache in the NSGA-III algorithm to avoid executing LDA configurations that had been previously executed.

# Chapter 6

# Contribution 3: Research Object Representation

As discussed in Section 2, most recommender systems use relatively simple retrieval models, such as *keyword matching* or the *Vector Space Model* (VSM) with basic *Term Frequency-Inverse Document Frequency* (TF-IDF) weighting. VSM is a spatial representation of text documents. In that model, each document is represented by a vector in an $n$-dimensional space, where each dimension corresponds to a term from the overall vocabulary of a given document collection.

In our case, we have defined a hierarchy of *spaces* to describe domains where *research objects* are described, processed and measured in different manners. The goal is modelling *research objects* in an efficient way while preserving the essential statistical relationships that are useful for tasks such as classification, summarization and similarity and relevance judgements.

## 6.1 Elements

Internally, our system uses the concept of *Research Objects* [38] to describe research resources. So, regardless of the input format we need to complete as much as possible its meta-information to store it following the principles of: unique ids, aggregation and annotation.

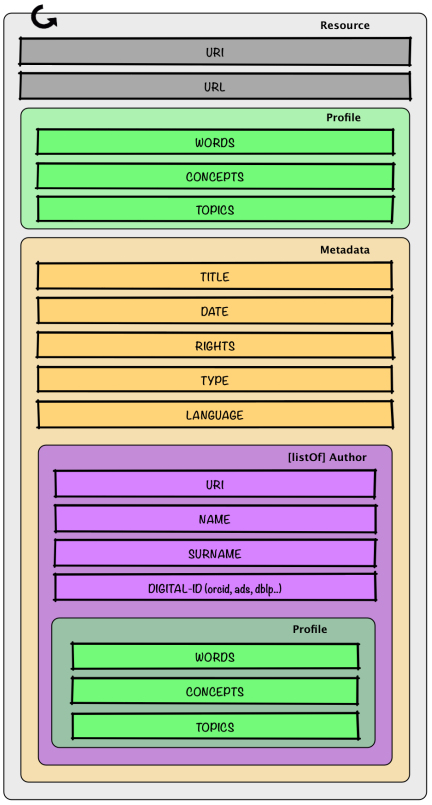



FIGURE 6.1: Internal resource representation

Taking into account the notation used by the *ro ontology* to describe resources [12], i.e. *Dublin Core Metadata Initiative (DCMI)*, we have defined *resource* (*Figure* 6.1) (*item* in recommender systems) as an entity identified by both a *Uniform Resource Identifier (URI)* and a *Uniform Resource Locator (URL)*, that is contextualized by *metadata*, described by a group of elements such as *words*, *concepts* and *topics* and may contain other *resources*.

*Metadata* is the conceptualization of all the meta-information associated to a resource such as *title*, *dates* (publication, creation), *license rights*, *format type* and the list of *authors*.

An *author* (*user* in recommender systems) is the primarily responsible entity for making the resource, and it is described by a *name* and a *surname* and identified by a *Uniform Resource Identifier (URI)* along with other digital identifications such as *orcid*, *ads id* and/or *dblp id*.

As shown above, this is not a typical recommender system where users and items are connected by rating or interest. Now, a user, i.e *author*, is related to an item, i.e *research*

*resource*, because she/he was one of its *creators*. As mentioned in Section 2, this system will use the *implicit feedback* technique to connect the elements of the domain.

## 6.2 Spaces

In the context of this work, we use the term *space* as a work domain characterized by the following aspects: First, a *dimension* defining its cardinality, i.e. number of elements. Second, a *collection of objects* that compose its population. Third, a *feature vector*, whose dimension is equal to the dimension of the space, which describes every resource as a chain of numerical values. Finally, a list of *operations* and *metrics* available to be executed over the objects.

All spaces are changing environments that should be adapted when new resources are added to the system but maintain a minimum set of constraints that make them different between them.

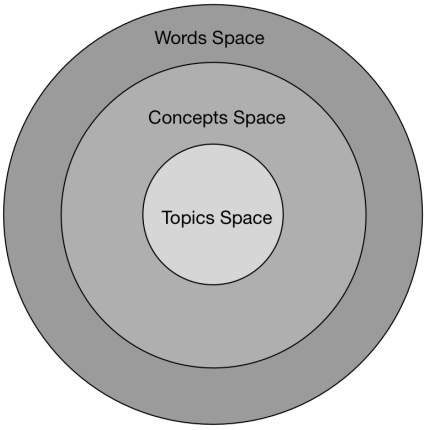



FIGURE 6.2: Domain Spaces

### 6.2.1 Words Space

The initial space and base for the rest of spaces is the *Words-Space.* It considers ROs as *regular-resources* described by words. A *regular-resource* is an entity identified by a *Uniform Resource Identifier (URI)* and a *Uniform Resource Locator (URL)*, which is described by meta-information such as title, date of publication, authors, license rights, format and so on. It has a *bag-of-words* describing its content and, as resource, it may aggregate other/s regular-resources.

The cardinality of this space is equal to the size of the vocabulary used to describe resources, i.e. number of distinct words, and the feature vector is composed with the frequencies of every word in the bag-of-words of the resource. Similar approaches [6] [72] [37] [43] use the *tf-idf* weighting schema to create the feature vectors, but in this case it is not adequate because the LDA model already considers the weight of a word based on its occurrences in different documents combining the Dirichlet distribution of words by topics and the Dirichlet distribution of topics by documents.

### 6.2.2 Concepts Space

The *Concepts-Space* is built over the *Words-Space*. This space is an extension of the *Word-Space* that also considers *concepts* to describe resources. Resources here are members of a more generic resource named *conceptual-resource*, that is an extended vision of regular-resources containing all the information described above along with a *bag-of-concepts* created from the *bag-of-words* of the resource. This conceptualization of words is a key task that can be specified by domain ontologies as showed in [73] or even handled by an identity function considering each word as a concept. This space is created to allow a new level of abstraction from the point of view of words.

Now, the vocabulary is composed by the list of distinct concepts used to describe all the *conceptual-resources* in the domain, so this space is a subspace of the *Words-Space* because its dimension, i.e cardinality of the vocabulary, is smaller or equal to it since a concept may describe more than one word. The feature vector, as in the previous space, contains the frequencies of each concept from the *bag-of-concepts* of the *conceptual-resource*.

### 6.2.3 Topics Space

Closing the hierarchy, the *Topics-Space* is built over the *Concepts-Space*, including *topic information* along with existing information for every resource. As described in section 2, considering a resource as a document in the LDA model, every resource has a topics distribution based on its *bag-of-concepts* and the *bag-of-concepts* of the rest of resources in the corpus. A *topic* is not a word nor a concept, but a Dirichlet distribution of concepts. Hence the number of topics may vary when corpus change. A resource, now

named *topical-resource*, contains a vector with the probability of each topic according to its content, i.e. its *bag-of-concepts*. This vector is the *topics distribution* of the *topical-resource*.

In this space, in addition to *topical-resources* there are also *author-profiles*. An *author-profile* is a temporal sorted chain of publications, i.e. research objects, along with the topics distribution of each resource. This information is really useful to make predictions and make recommendations as will be showed in section 8.

The cardinality of this space is equal to the number of topics, being these topics the vocabulary of the space. The feature vector is the topics distribution of each *topical-resource* as previously mentioned. This space is where the system works to make inferences and analysis.

## 6.3 Evaluation

First, the *Hoarder* application downloaded records (metainformation and resources) from 10 different OAI-PMH data providers as indicated in Section 4.2.3. After that, the *Harvester* application composed research object from them as mentioned in Section 4.3.1. Now, using these research objects serialized as JSON in textual files, the RESSIST application create *regular resources* from them, then *conceptual resources* and, finally, *topical resources*.

These elements will be processed by Spark, so we have implemented a *Spark Function* to serialize a *ResearchObject* as a *RegularResource*. This is an easy task because the *ResearchObject* serialized as JSON contains all the information (even more) that a *RegularResource* needs. This function is as follows:

```
1  @Component
2  public class RRParser implements Function<ResearchObject, RegularResource> {
3
4      @Override
5      public RegularResource call(ResearchObject researchObject) throws Exception {
6          String uri                      = researchObject.getUri();
7          String url                      = researchObject.getUrl();
8          String title                    = researchObject.getMetainformation().getTitle();
9          String published                = researchObject.getMetainformation().getPublished();
10         Buffer<String> bagOfWords       = JavaConverters.asScalaBufferConverter(researchObject
       .getBagOfWords()).asScala();
11         Buffer<RegularResource> innerResources   = JavaConverters.asScalaBufferConverter(new
       ArrayList<RegularResource>()).asScala();
12         List<Author> authorsList = new ArrayList<Author>();
13         for (Creator creator: researchObject.getMetainformation().getCreators()){
```

```
14              String authorUri    = creator.getUri();
15              String name         = creator.getName();
16              String surname      = creator.getSurname();
17              authorsList.add(new Author(authorUri,name,surname));
18          }
19          Buffer<Author> authors = JavaConverters.asScalaBufferConverter(authorsList).asScala();
20          Metadata metadata = new Metadata(title, published,authors);
21          RegularResource regularResource = new RegularResource(uri,url,metadata,bagOfWords,
        innerResources);
22          return regularResource;
23      }
24  }
```

LISTING 6.1: *RegularResource* conversion function from a *ResearchObject*

We have included some *scala.JavaConverters* in our RESSIST (*Java-based*) application to use the metrics that we have implemented in Scala language. We used Scala to define metrics, calculus and correlations because Spark is implemented in Scala and then, a translation from java to scala during the LDA processing is not needed. However the RESSIST application is implemented in Java to manage the global process reusing some of the most useful java frameworks such as Spring-Framework, Apache-Camel, JSW, and so on.

Once the conversion from *ResearchObject* to *RegularResource* is completed, then a *Words-Space* is created. At this point, the vocabulary of the space is composed by the distinct words from the *bagOfWords* of all *RegularResources*.

Then, a new conversion to *ConceptualResource* is needed to reach the *Concepts-Space*. Actually, this conversion is direct being the same word a concept. Future works will try to obtain general concepts from words but in a sensitive way because they cannot be as general as possible because the LDA model will not detect useful topics. We will need to define a balanced criteria to obtain concepts from words in a useful way.

```
1  @Component
2  public class CRParser implements Function<RegularResource, ConceptualResource> {
3
4      @Override
5      public ConceptualResource call(RegularResource regularResource) throws Exception {
6          return new ConceptualResource(regularResource);
7      }
8  }
```

LISTING 6.2: *ConceptualResource* conversion function from a *RegularResource*

Once we have *ConceptualResources*, the vocabulary is now composed by concepts instead of words. Finally, a *Topics-Space* is built using these *ConceptualResources*. Now is not possible a direct conversion, firs of all a new LDA Model is created in this space

defining a set of topics in the corpus and generating new topics distribution for each *ConceptualResource*, then the *TopicalResources* are created.

This space, and this LDA model, are used to make predictions and recommendations. Now the vocabulary is composed by topics (probability distributions), and the feature vectors are based on these distributions for each resource.

# Chapter 7

# Contribution 4: Research Object Similarity Measure

As stated previously, the system will make predictions, inferences and recommendations based on research objects and any other useful information derived from them. For this, some metrics are required, mainly similarity measures, to connect resources, authors and extract knowledge from these relationships.

Measuring the similarity of ROs is a key task from which to obtain useful knowledge. Its definition must be general enough to overcome the particular characteristics of the different types of resources that ROs aggregate. Thus, similarity evaluations may show differences on the accuracy between *regular-resources*, *conceptual-resources* or *topical-resources* but not between different types of content in the same resource expression, i.e. a textual-based *regular-resource* and an image-based *regular-resource*.

Because an RO contains two kinds of information: *context-based*, i.e. authors, license rights, format..., and *content-based*, i.e. text, image, code.., the similarity metric includes both concepts. In short, the measure to calculate the similarity between two ROs is a weighted sum of *context-based* similarity ($sim_{ctx}$) and *content-based* similarity ($sim_{cont}$):

$$sim(R_i, R_j) = \alpha * sim_{cont}(R_i, R_j) + (1 - \alpha) * sim_{ctx}(R_i, R_j) \tag{7.1}$$

where $\alpha \in [0, 1]$

Depending on the nature of the resource, i.e. *regular-resource*, *conceptual-resource* or *topical-resource*, each of these *content-based* and *context-based* similarity measures are different as detailed above.

## 7.1 Content-based Similarity

### 7.1.1 Under Frequency

Both *Words-Space* and *Concepts-Space* are based on frequency vectors as feature vectors. The first space counts word frequencies and the second one counts concept frequencies. The expression that describes the content similarity between resources based on frequency vector is the same in both spaces.

As mentioned before in Section 6, a *regular-resource* is an entity that may aggregate other/s *regular-resource/s* and contains identification and descriptive information such as title, authors and a bag-of-words describing its content. Regardless of whether it is a textual resource or an image resource or any other, it is annotated by a list of words to describe what the content means. Future works will take into account the type of the content to make a more specific similarity metric, but at the present time the *content-based* similarity measure considers that group of words to measure how similar two resources are.

Because a *regular-resource* may contain aggregated resources, the feature vector used to measure the similarity is the vectorial sum of the feature vectors of each nested resource. We used the *cosine similarity* based on the *Euclidean dot product* as similarity measure to take into account the proportional use of words instead of frequency values directly. So, the ***content-based similarity measure*** is:

$$sim_{cont}(R_i, R_j) = cos(\hat{r_i}, \hat{r_j}) \tag{7.2}$$

where $\hat{r_i}$ is the feature vector of the research object $R_i$ described as *regular-resource* or *conceptual-resource* and $cos(\hat{r_i}, \hat{r_j})$ is the *cosine similarity*:

$$cos(P, Q) = cos(\theta) = \frac{P \cdot Q}{\|P\|\|Q\|} = \frac{\sum_{i=1}^{n} P_i \times Q_i}{\sqrt{\sum_{i=1}^{n} (P_i)^2} \times \sqrt{\sum_{i=1}^{n} (Q_i)^2}} \tag{7.3}$$

### 7.1.2 Under Topics Distribution

However, in the *Topics-Space* the feature vector is a topics distribution expressed as vector of probabilities. As showed in Section 6, when a resource is aggregated by other resources, the *bag-of-concepts* used to calculate the topics distributions is the sum of concepts used in each nested resource. For this reason, the topics distribution of the root of an aggregation is enough to describe the complete research object.

Taking into account this premise, the similarity measure between two *topical-resources* will be based on the distance between their topics distributions. Since they are Dirichlet distributions (probability mass functions), the measure used was the *Jensen-Shannon divergence*, which can be defined as the average of the *Kullback-Leibler (KL) divergence* between them. KL has two major problems: in the case that one of topics distribution is zero, KL is not defined and it is not symmetric, what does not fit well with semantic similarity measures which in general are symmetric [62]. To solve these problems, *Jensen-Shannon divergence* considers the average of the distributions as below [20]:

$$JSD(p, q) = \sum_{i=1}^{T} p_i * \log \frac{2 * p_i}{p_i + q_i} + \sum_{i=1}^{T} q_i * \log \frac{2 * q_i}{q_i + p_i} \tag{7.4}$$

where $T$ is the number of topics and $p, q$ are the topics distributions

Our ***content-based similarity measure*** use the *Jensen-Shannon divergence* transformed into a similarity measure as follows [22] :

$$sim_{cont}(R_i, R_j) = 10^{-JSD(p,q)} \tag{7.5}$$

where $R_i, R_j$ are the research objects and $p, q$ the topics distributions of each of the *topical-resources* describing them.

## 7.2 Context-based Similarity

Currently, context-based similarity is only related to author-based similarity. The plan is to include more elements in the future, both extracted directly from the resource or inferred, so as to increase the accuracy of the measure. But now the system must work properly with authors:

$$sim_{ctx}(R_i, R_j) = sim_{authors}(R_i, R_j) \tag{7.6}$$

Before obtaining the similarity of authors we need to define how an author is described. Similar to feature vectors to describe the content of a resource, an author is represented by a vector that describes adequately his/her most relevant aspects to allow us to take measures between them. The dimension of this vector will depend on the cardinality of the used space. As in the case of resources, two types of feature vectors exist: frequency-based and topics-based.

### 7.2.1 Under Frequency

This similarity will be used both in *Words-Space* and *Concepts-Space* because it considers that an author is described by each of the feature vectors of the research objects published by him/her. Recently, a temporal combination has been proposed to obtain a valid similarity measure between authors [56] . They defined an *author similarity (AS)* based on *cosine similarity* of the feature vectors for a given interval time:

$$AS_{cos}(A, B, t_1, t_2) = cos(\sum_{i=t_1}^{t_2} \hat{a_i}, \sum_{i=t_1}^{t_2} \hat{b_i}) \tag{7.7}$$

where $\hat{a_i}$ and $\hat{b_i}$ are the feature vectors of the authors $A$ and $B$ in the i-th year.

Authors usually publish more than one research object in the same year, so the feature vector for that i-th year will be the vectorial sum of feature vectors of each research object published.

However, this metric does not take into account possible common shifts of interests of the authors. In fact, if the author $A$ worked on topic $T_1$ and then shifted to topic $T_2$,

he will be considered similar to author $B$ who was originally in $T_2$ and then moved to $T_1$. To avoid this problem, a metric that pays attention to the period of time in which an author addresses a specific topic is needed, rewarding common trajectories. Hence, in order to strengthen the importance of the time factor, a partial similarity recursively on increasingly shorter time intervals is proposed and the final similarity is the average of the results. More formally, a *temporal author similarity (TAS)* between an author $A$ and an author $B$ in the interval $t_1 - t_2$ is:

$$TAS(A, B, t_1, t_2) = \frac{\sum_{i=0}^{m} \left[ (\sum_{j=0}^{2^i-1} AS(A, B, t_1 + \lceil \frac{j \cdot (t_2 - t_1)}{2^i} \rceil, t_1 + \lfloor \frac{(j+1) \dot{(} t_2 - t_1)}{2^i} )) / 2^i \right]}{m+1} \tag{7.8}$$

where $m = \lfloor log(t_2 - t_1) \rfloor$

This temporal author similarity covers well the case in which both authors are present in the same time interval, however an author may have no publications in some of the years inside the interval. Then a penalty $P$ is applied as the average of $AS$ of $n$ authors randomly extracted from the input. In our opinion, this penalty should be changed by the feature vector of the last publication, then the intervals of time without publications would have the same feature vector than the last year with publication. Thus, our similarity measure between two authors is:

$$sim_{author}(A, B) = TAS(A, B, t1, t2) \tag{7.9}$$

where $t_1$ is the oldest publication date of both authors and $t_2$ the newest one.

Once we know the similarity measure between two authors, we can calculate the ***author-based similarity measure*** between ROs as the minimum similarity value between the authors of each research object:

$$sim_{authors}(R_i, R_j) = min(sim_{author}(a_{im}, a_{jn})) = min(TAS(A, B, t1, t2)) \tag{7.10}$$

where $a_{im}$ is the m-th author of the *regular — conceptual-resource* $i$, and $a_{jn}$ the n-th author of the *regular — conceptual-resource* $j$.

The feature vectors used in this calculus contain frequencies of words in *Words-Space* and frequencies of concepts in *Concepts-Space*.

### 7.2.2 Under Topics Distribution

Now the feature vector is a multinomial probability distribution so the challenge here is to produce a consensus topics distribution for each author by combining appropriately the topics distributions of their publications. The most popular choice for this aggregation is *Linear Pooling*, which assigns each individual forecast a weight which reflects the importance of the publication, but if we provide an equal weight to every probability the method reduces to an arithmetic average. A *Generalized Linear Pooling* extends the previous approach considering the possibility of negative weights. However [59] any linear combination of (calibrated) forecast is uncalibrated and lacks sharpness then a *Beta-transformed Linear Pooling* is proposed applying a *Beta transformation* to linear pooling operators in order to add a recalibration step to the process and improve their performance. A probability $P_G(A)$ is said to be calibrated if $P(Y_k|P_G(A_k)) = P_G(A_k, k = 1 \dots K)$ [59]. Sharpness refers to the concentration of the aggregated distribution. The more concentrated it is, the sharper it is.

Intuitively, aggregation operators based on multiplication seem more appropriate than those based on addition. *Log-linear Pooling* is a linear operator of the logarithms of the probabilities that does not preserve independence and does not verify the marginalization property.*Generalized Logarithmic Pooling* extends it by adding an arbitrary bounded function. On the other hand, instead of establishing a pooling formula from an axiomatic point of view, the aggregation of two distributions could be those that share properties (moments or conditional probabilities) and minimize the KL divergence between them.

Furthermore, as showed in several simulation studies [3], *linear pooling* performs poorly relative to other pooling formulas with a multiplicative instead of an additive structure. Also, many of non-linear methods involve a large number of parameters, making them computationally complex and susceptible to over-fitting. By contrast, parameter-free approaches, such as the median or the geometric mean of the odds, are too simple to be able to incorporate the use of training data optimally.

Recently, an approach based on the *log-odds statistical model* of the data has been proposed [65] being an alternative way to express probabilities using the odds ratio.

However, the LDA model considers topics distributions as Dirichlet distribution, i.e continuous multivariate probability distributions , then we can combine them to get a more general topics distribution using the Bayes' Theorem. Thus, considering the following topics distributions $(td_1, td_2)$ for the research objects $(R_1, R_2)$ and the topics $(T_1, T_2, T_3)$:

$$td_1 = (t_{11}, t_{12}, t_{13})$$
$$td_2 = (t_{21}, t_{22}, t_{23})$$

and taking into account that:

$$t_{ij} = p(T_i/R_j)$$

the consensus topics distribution $td_f$ will be:

$$td_f = (P(T_1/R_1, R_2), P(T_2/R_1, R_2), P(T_3/R_1, R_2))$$

As $R_1$ and $R_2$ are independent and using the Bayes' theorem we get:

$$P(T_i/R_1, R_2) = \frac{P(R_1) \cdot P(R_2)}{P(R_1, R_2)} \times \frac{P(T_i/R_1) \cdot P(T_i/R_2)}{P(T_i)} = \alpha \times \frac{P(T_i/R_1) \cdot P(T_i/R_2)}{P(T_i)} \tag{7.11}$$

where $\alpha$ is a class-independent term depending only on the data. As we have measured these data, its value is not interesting here (we are not doing model comparisons), so we treat it as a normalization constant which ensures the Dirichlet constraint that $\sum_k P(T_k/R_1, R_2) = 1$.

Now that we know how to combine topic distributions, we can redefine the *author similarity (AS)* expression using the *Jensen-Shannon Divergence* as a distance measure of topics distributions and taking its similarity expression for a given interval time:

$$AS_{JSD}(A, B, t_1, t_2) = 10^{-JSD(\hat{a_{12}}, \hat{b_{12}})} \tag{7.12}$$

where $\hat{a_{12}}$ and $\hat{b_{12}}$ are the consensus topics distributions of authors $A$ and $B$ for the interval of time $t_1 - t_2$.

## 7.3 Evaluation

At this point, we have a *Topics-Space* composed by *TopicalResources*, and created from the *ConceptualResources* that were directly generated from the *RegularResources* that describe the research objects of the corpus. In this space, a LDA model is generated to create the *TopicalResources* that contain *ConceptualResources* along with topics distributions. Using these distributions, the RESSIST application can calculate the similarity measures based on topics between a couple of resources, i.e. between two research objects.

Applying the configuration of the LDA model suggested by the learning algorithm detailed in Section 5.2, i.e. 12 topics, alpha=6.1 and beta=1.1, our application builds a Topic Model that defines probabilistic distributions for each resource based on its *bag-of-concepts* (in fact, as before mentioned, based on words). After a stemming process, the concepts are reduced to their stem, i.e. base or root. Taking the list of stems for each resource and their frequencies, the system build 12 topic that contain, in a different proportion, the list of stems of the corpus. This distribution of *stems* by topics is listed in the table 7.1, showing only the 20 most relevant stem for each topic.

Using these topics, i.e probabilistic distributions of stems (from concepts/words), the model assigns a distribution of topics for each resource, table 7.3, then our recommender system creates a similarity matrix calculating the similarity measurements between all the research objects of the corpus. The table 7.2, for example, shows a column of that matrix. It contains the similarity measurements between the referenced research object, with a similarity equal to 1, and the rest of research objects of the corpus. In the table we also identify the data provider where the resource was published to identify the research area where the publisher, in this case *Innovare Journal*, has classified the resource: Agricultural Science (IJAGS), Business Management (IJBM), Education (IJOE), Ayurvedic Science (IJAS), Engineering and Technology (IJET), Health Science (IJHS), Life Science (IJLS), Medical Science (IJMS), Social Science (IJSS) and Science

| Topic | Most Frequent Terms |
|---|---|
| 0 | 'collect' ,'extract' ,'procedur' ,'evalu' ,'univers' ,'plant' ,'chemic' ,'health' ,'medicin' ,'found' ,'standard' ,'research' ,'antimicrobi' ,'revis' ,'receiv' ,'pharmaci' ,'abstract' ,'keyword' ,'accept' ,'screen' |
| 1 | 'found' ,'review' ,'email' ,'scienc' ,'articl' ,'accord' ,'receiv' ,'keyword' ,'abstract' ,'revis' ,'accept' ,'import' ,'which' ,'other' ,'introduct' ,'refer' ,'gmail' ,'studi' ,'innovar' ,'journal' |
| 2 | 'afford' ,'itself' ,'arrang' ,'lesson' ,'mobil' ,'integr' ,'citizen' ,'reach' ,'strong' ,'charg' ,'allow' ,'equip' ,'altern' ,'opportun' ,'start' ,'provis' ,'build' ,'offer' ,'challeng' ,'subject' |
| 3 | 'defin' ,'first' ,'review' ,'receiv' ,'abstract' ,'keyword' ,'articl' ,'revis' ,'accept' ,'other' ,'point' ,'which' ,'refer' ,'group' ,'journal' ,'innovar' ,'paper' ,'inform' ,'anoth' ,'conclus' |
| 4 | 'email' ,'where' ,'articl' ,'revis' ,'keyword' ,'abstract' ,'accept' ,'which' ,'introduct' ,'refer' ,'innovar' ,'journal' ,'engin' ,'through' ,'gener' ,'variou' ,'conclus' ,'receiv' ,'consid' ,'techniqu' |
| 5 | 'method' ,'should' ,'receiv' ,'abstract' ,'keyword' ,'revis' ,'accept' ,'research' ,'about' ,'effect' ,'introduct' ,'which' ,'studi' ,'refer' ,'perform' ,'achiev' ,'journal' ,'innovar' ,'aspect' ,'materi' |
| 6 | 'develop' ,'signific' ,'other' ,'effect' ,'should' ,'through' ,'introduct' ,'refer' ,'which' ,'receiv' ,'abstract' ,'keyword' ,'increas' ,'accept' ,'requir' ,'innovar' ,'journal' ,'articl' ,'revis' ,'total' |
| 7 | 'agent' ,'treatment' ,'therefor' ,'further' ,'method' ,'result' ,'present' ,'system' ,'scienc' ,'articl' ,'differ' ,'revis' ,'receiv' ,'keyword' ,'abstract' ,'respect' ,'accept' ,'effect' ,'activ' ,'absorb' |
| 8 | 'colleg' ,'patient' ,'treatment' ,'email' ,'result' ,'articl' ,'research' ,'receiv' ,'abstract' ,'keyword' ,'revis' ,'accept' ,'object' ,'indor' ,'qualiti' ,'ayurveda' ,'group' ,'manag' ,'increas' ,'method' |
| 9 | 'gmail' ,'email' ,'receiv' ,'revis' ,'accept' ,'studi' ,'journal' ,'innovar' ,'district' ,'rural' ,'articl' ,'abstract' ,'keyword' ,'research' ,'primari' ,'patient' ,'medic' ,'occur' ,'intern' ,'adult' |
| 10 | 'appli' ,'respect' ,'nation' ,'evalu' ,'second' ,'countri' ,'properti' ,'intern' ,'univers' ,'those' ,'sampl' ,'resist' ,'express' ,'howev' ,'includ' ,'process' ,'function' ,'number' ,'basic' ,'avail' |
| 11 | 'discuss' ,'method' ,'result' ,'receiv' ,'abstract' ,'keyword' ,'accept' ,'found' ,'introduct' ,'studi' ,'refer' ,'scienc' ,'journal' ,'innovar' ,'articl' ,'email' ,'differ' ,'revis' ,'under' ,'india' |

TABLE 7.1: Distribution of terms by topics

| Resource | Similarity | Research Object | Provider |
|---|---|---|---|
| 1 | 1.0 | SIMULTANEOUS ESTIMATION OF IRBESARTAN AND ATORVASTATIN BY FIRST ORDER DERIVATIVE SPECTROSCOPIC METHOD IN THEIR SYNTHETIC MIXTURE USE IN HYPERTENSION CONDITION | IJS |
| 2 | 0.9985002911336149 | DEVELOPMENT AND VALIDATION OF ANALYTICAL METHOD FOR IRBESARTAN AND ATORVASTATIN BY SIMULTANEOUS EQUTION SPECTROSCOPIC METHOD | IJS |
| 3 | 0.9984274995642390 | SIMULTANEOUS ESTIMATION OF IRBESARTAN AND ATORVASTATIN BY Q ABSORPTION RATIO METHOD IN THEIR SYNTHETIC MIXTURE USE IN CARDIAC CONDITION | IJS |
| 4 | 0.9950251177695804 | ARTEMETHER LUMEFANTRINE LOADED LIPOSPHERES EVALUATION OF PROPERTIES OF SOLUTOL HS 15 AND SOLUPLUS ON THE IN VITRO PROPERTIES | IJS |
| 5 | 0.9890510173748973 | EFFECT OF LYOPHILIZATION ON THE PHYSICOCHEMICAL AND PHYSICOTECHNICAL PROPERTIES OF ASPIRIN-LOADED LIPOSPHERES | IJS |
| 6 | 0.9812091925351933 | COMPATIBILITY OF BEAUVERIA BASSIANA (BALS.) VUILL ISOLATES WITH SELECTED INSECTICIDES AND FUNGICIDES AT AGRICULTURE SPRAY TANK DOSE | IJAGS |
| 7 | 0.9487017007511238 | DIFFERENT MODELS TO EVALUATE ANTIMICROBIAL AGENTS-A REVIEW | IJLS |
| 8 | 0.9365837211037017 | PREPARATION OF CHITOSAN STABILIZED OFLOXACIN- GOLD NANO CONJUGATE FOR THE IMPROVED ANTI BACTERIAL ACTIVITY AGAINST HUMAN PATHOGENIC BACTERIA | IJMS |
| 9 | 0.8089238194923263 | An overall review on Obesity and its related disorders | IJLS |
| 10 | 0.21158453507688826 | RESEARCH ON FORMULATION AND EVALUATION OF INSITU MUCOADHESIVE NASAL GELS OF METOCLOPRAMIDE HYDROCHLORIDE | IJMS |
| .. | .. | .. | ... |

TABLE 7.2: Similarity measures between research objects from the same data provider.

(IJS). This classification is used as reference during the tests to check the validity of the results.

A first important behaviour showed in the table 7.2 is that, if considering as similar only the research objects with a similarity value greater than 0.5, only 8 research objects are similar to the referenced one, that is only the 8% of the corpus. This exhaustive classification, in our opinion, is caused by the high number of topics, 12, and the low value of *alpha*, 6.1. With these values the research objects have a low concentration of topics

and then different research objects can be described by *strong* topic distributions, i.e high values for some topics and low for the rest, avoiding middle values. In fact, the difference between the 8th research object, *"An overall review on Obesity and its related disorders"*, and the 9th, *"RESEARCH ON FORMULATION AND EVALUATION OF INSITU MUCOADHESIVE NASAL GELS OF METOCLOPRAMIDE HYDROCHLORIDE"*, is almost 0.6 points. This behaviour is common for the rest of columns of the similarity matrix, as showed also in the table 7.4.

In addition, as expected, the most similar research objects are published in the same domain, IJS, so they belong to the same research area, Science. However, research objects from other research areas such as IJAGS (Agricultural Science), IJLS (Life Science) and IJMS(Medical Science) are also present in the column as similar researches. The reason is because our similarity measure does not handle the meaning of paragraphs or the relevance of terms considered by an author, it takes into account the frequency of terms to build a model based on their probabilities to belong to a topic, i.e. a cluster, and RESSIST uses these probabilities to make relationships between them based on their content and their contextual information. In our opinion this is useful, as showed later in Section 8, to discover relationships between researches from different domains, from different research areas, from different authors. For example in the table 7.2, the system detects that the research *"SIMULTANEOUS ESTIMATION OF IRBESARTAN AND ATORVASTATIN BY FIRST ORDER DERIVATIVE SPECTROSCOPIC METHOD IN THEIR SYNTHETIC MIXTURE USE IN HYPERTENSION CONDITION"* about Science is similar, with a similarity measure equals to 0.981209, to the research *"COMPATIBILITY OF BEAUVERIA BASSIANA (BALS.) VUILL ISOLATES WITH SELECTED INSECTICIDES AND FUNGICIDES AT AGRICULTURE SPRAY TANK DOSE"* about Agricultural Science because both publications contain, in a similar way, the set of terms described by the topics in the table 7.1 , i.e. the words more frequently used are listed in topics 7, 11, 6 and 4, and the less frequently used are listed in topics 2, 9 and 10.

So, as showed in table 7.3, the topic 7 is the most representative for the research object *"SIMULTANEOUS ESTIMATION OF IRBESARTAN AND ATORVASTATIN BY FIRST ORDER DERIVATIVE SPECTROSCOPIC METHOD IN THEIR SYNTHETIC MIXTURE USE IN HYPERTENSION CONDITION "* , as well as the topics 11, 6 and 4. Then, research objects following a similar distribution of topics are more

| R | S | T0 | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 | T11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ijs | 0.0045 | 0.0049 | 0.0028 | 0.0044 | 0.0057 | 0.0044 | 0.0060 | 0.9459 | 0.0054 | 0.0034 | 0.0038 | 0.0081 |
| 2 | ijs | 0.0053 | 0.0056 | 0.0032 | 0.0051 | 0.0075 | 0.0052 | 0.0075 | 0.9342 | 0.0070 | 0.0038 | 0.0047 | 0.0105 |
| 3 | ijs | 0.0055 | 0.0060 | 0.0033 | 0.0051 | 0.0075 | 0.0049 | 0.0079 | 0.9338 | 0.0065 | 0.0039 | 0.0047 | 0.0103 |
| 4 | ijs | 0.0061 | 0.0061 | 0.0044 | 0.0060 | 0.0098 | 0.0053 | 0.0107 | 0.9259 | 0.0053 | 0.0041 | 0.0058 | 0.0098 |
| 5 | ijs | 0.0058 | 0.0070 | 0.0043 | 0.0071 | 0.0135 | 0.0059 | 0.0115 | 0.9138 | 0.0057 | 0.0045 | 0.0071 | 0.0131 |
| 6 | ijags | 0.0079 | 0.0095 | 0.0053 | 0.0092 | 0.0117 | 0.0079 | 0.0128 | 0.8981 | 0.0083 | 0.0064 | 0.0082 | 0.0141 |
| 7 | ijls | 0.0091 | 0.0132 | 0.0075 | 0.0139 | 0.0275 | 0.0097 | 0.0163 | 0.8621 | 0.0093 | 0.0091 | 0.0079 | 0.0140 |
| 8 | ijms | 0.0163 | 0.0113 | 0.0066 | 0.0081 | 0.0279 | 0.0102 | 0.0133 | 0.8495 | 0.0089 | 0.0082 | 0.0092 | 0.0299 |
| 9 | ijls | 0.0102 | 0.0238 | 0.0086 | 0.0211 | 0.0233 | 0.0334 | 0.0466 | 0.7382 | 0.0460 | 0.0170 | 0.0144 | 0.0168 |
| 10 | ijms | 0.0161 | 0.0192 | 0.0102 | 0.0149 | 0.0314 | 0.0156 | 0.0179 | 0.2156 | 0.0198 | 0.0123 | 0.0114 | 0.6151 |

TABLE 7.3: Distribution by topics of research objects listed in table 7.2

| Similarity | Research Object | Provider |
|---|---|---|
| 1.0 | RIVIEW OF SHRINGA , ALABY AND CUPPOING THERAPY | IJAS |
| 0.9901248961071275 | PREVALENCE, ETIOLOGY AND CLINICAL FEATURES OF SKELETAL FLUOROSIS: A CRITICAL REVIEW. | IJMS |
| 0.9898432754651388 | ALL ABOUT YOGA | IJHS |
| 0.9783233516492001 | A CASE STUDY OF GIFTED CHILD. | IJOE |
| 0.9752164237190362 | ROLE OF AN IMPORTANCE OF ACTIVITIES IN SCHOOL ENVIRONMENT. | IJOE |
| 0.9733055901744754 | A STUDY OF EMOTIONAL INTELLIGENCE OF HIGHER SECONDARY SCHOOL TEACHERS OF MADHYA PRADESH | IJOE |
| 0.9716959991552354 | Relationship between cigarette smoking and body mass index in the Italian population | IJHS |
| 0.9671450145579238 | IMPACT OF ACTIVE LEARNING STRATEGIES TO ENHANCE STUDENT PERFORMANCE | IJOE |
| 0.9474338200755947 | DO LEADERSHIP QUALITIES DETERMINE COMPETENT PRINCIPALS | IJOE |
| 0.9246011509355332 | INFLUENCE OF ELECTRONIC MEDIA ON CHILDREN'S PERSONALITY DEVELOPMENT | IJSS |
| 0.9154436939927828 | AN INTRODUCTION OF PROBLEM BASED LEARNING IN IMS, BHU | IJMS |
| 0.9064540067697475 | An overall review on Obesity and its related disorders | IJLS |
| 0.8396635494170802 | KNOWLEDGE OF MEDICAL NEGLIGENCE AMONG MEDICAL STUDENTS | IJMS |
| 0.27918402806682086 | AYURVEDA AND MENTAL HEALTH | IJAS |
| .. | .. | ... |

TABLE 7.4: Similarity measures between research objects from different data providers.

similar to it than the rest. This explains why a research object about Agricultural Science is more similar to a research object about Science than even other research objects also classified in Science. Depending on the stemming process, these similarities may vary, so that is a key task in our system. At this moment, we have used the Lucene classifier as the stemming algorithm, but in future work we will develop some variations to improve the accuracy of our classification procedure.

Moreover, as the table 7.4 shows, the system is not influenced by the type of the data provider used to collect the research objects, i.e. by the research area where a research object is classified. The system has detected two research objects that are the same research object, *"CLINICAL-COMPARATIVE STUDY OF VIRECHAN & PAKSHAGHATARI GUGGULU ON PAKSHAGHAT W.R.S. TO HEMPIPLIGIA"*, but that they were published in two different data providers: IJLS and IJAS. This multiple classification express that we have considered in our model, a research object may be oriented to more than one topic.

It is important to mention that the research area where a research object is focused may

be different from the content of the research object. The table 7.4 shows also a research object titled *"RIVIEW OF SHRINGA , ALABY AND CUPPOING THERAPY"* published in the *Innovare Journal of Ayurvedic Science* (IJAS) data provider that is more similar to other research object published in different data providers such as the *Innovare Journal of Medical Science* (IJMS) data provider, the *Innovare Journal of Health Science* (IJHS) or even the *Innovare Journal of Education* (IJOE). As previously mentioned, this occurs because our similarity measure is only based on the content (terms) and the authors of the research objects, instead of the research area or keywords. Future works will include a more complex semantic analysis to compose a similarity measure based not only in words or concepts, but also in the meaning of paragraphs or even in the ideas included in the conclusion section, for example.

The figure 7.1 shows the graph created from the similarity matrix. Some clusters appear, as previously mentioned, because a high number of topics have been defined in the model.

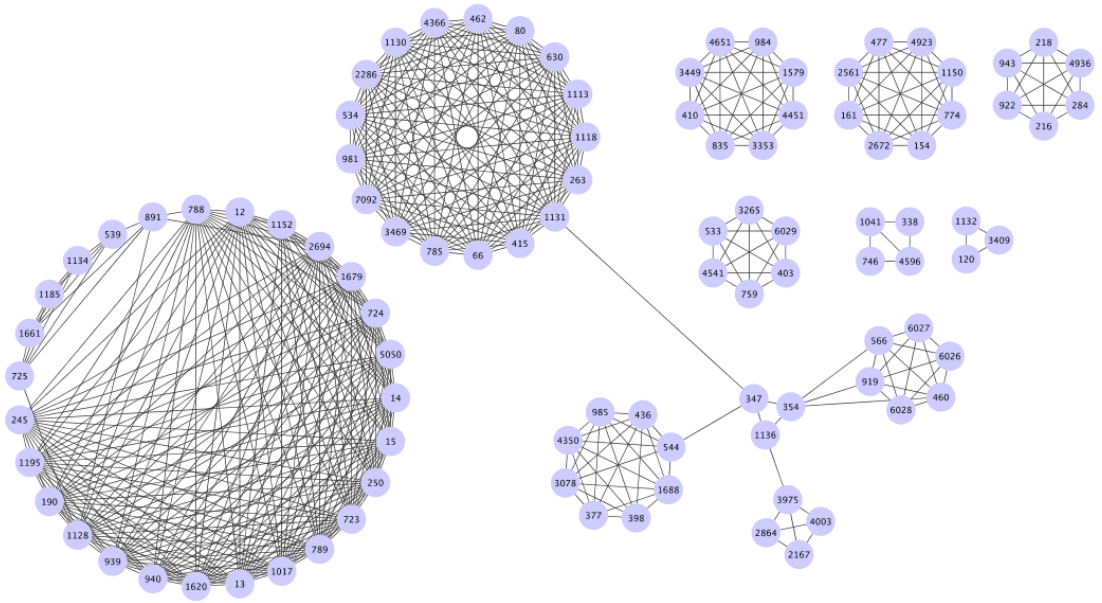



FIGURE 7.1: Research Object-Graph built from the test corpus (Appendix D)

# Chapter 8

# Contribution 5: Design of Research Objects-based Recommendations

At this point, our system is able to make recommendations based on research objects using the similarity measure defined in Section 7. In any recommender system, *users* and *items* are usually the main actors. In our case users are *authors* and items are *research objects*.

First of all, it is important to mention that although a research object can be added to the system at any time, the system processes all resources in a periodically scheduled batch mode. This way of working allows considering all research objects when the new LDA topic model is created. A resource is processed when a topic distribution is assigned to its content and its list of authors have also updated their profiles according to the new topic distribution of the content of their publications. Using this information, the system is able to calculate the similarity between all resources, creating a *similarity-matrix*. In fact, not only one matrix is created, but also a *research object-based* matrix and an *author-based* matrix. The first one contains *research object-based*, *content-based* and *context-based* similarity measurements between research objects, while the second one contains *author-based* similarity measurements between the authors. These matrices can be seen as graphs connecting research objects as well as authors.

We then propose some designs of recommendations based on these matrices with the aim of providing authors support for their researches. They tried to extract useful knowledge from the research objects and their relationships.

## 8.1 Recommendation 1: Route-of-Knowledge

**Goal**: Find a list of ROs connecting two research areas

An author having knowledge about some area, for instance Computer Graphics, wants to explore another area, for instance Astrophysics in order to identify papers or ROs that may be relevant to him/her. From a classical point of view, the next step for the author would be to read papers and/or books about Astrophysics to gain more knowledge about this area but this could be quite hard. Our recommender system offers a soft transition between these areas showing a sorted list of the most representative ROs that connect them, allowing authors to acquire the final knowledge in a incremental way.

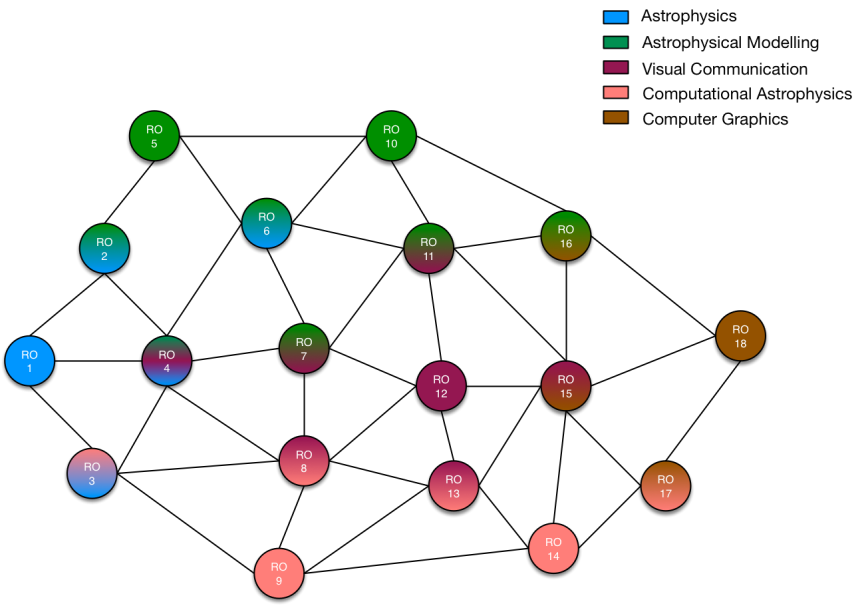



FIGURE 8.1: Research Object-Graph sample

The system builds an undirected graph to model pairwise similarity relations between ROs named *ro-graph*. It is based on the previously mentioned similarity measures (*Section* 7). First, the author types words or concepts describing the starting and the ending research areas. Then, the system obtains the topics distributions where these words or concepts have higher probabilities. For each of these topics distributions, the system

chooses the most similar ROs based on the distance measure defined in Section 7.4. Now, using the Dijkstra's algorithm [31], the system will obtain the minimum cost path between these ROs in the *ro-graph*. The reason to use this algorithm is because it is one of the most well-know algorithms to solve the shortest path problem (SPP) with acceptable performance. The route will be presented to the author as a sorted list of ROs to gain the desired knowledge.

Continuing with the example, the author will receive a sorted list of ROs to pass from Computer Graphics to Astrophysics, being the first RO the most representative research object in Computer Graphics (based on words/concepts provided by the author) and the last one the most representative research object in Astrophysics. The rest of ROs in the chain depend on the similarity measure used to connect the ROs in the graph. Therefore, an author may choose a *direct*, a *uniform*, or a *balanced route-of-knowledge*.

- ***direct***: Based exclusively on the *content* of ROs, this solution takes into account the content-based similarity measure under topics distribution described in Section 7.1.2 to only connect the ROs of similar *research areas*. For example, $RO_1$, $RO_3$, $RO_9$, $RO_{14}$, $RO_{17}$ and $RO_{18}$ (*Figure* 8.1).
- ***uniform***: Based exclusively on the *context* of ROs, this solution takes into account the content-based similarity measure under topics distribution described in Section 7.2.2 to only connect the ROs built in a *similar way*. At present, the system prioritizes ROs with similar authors, but future work will incorporate other aspects such as style of writing, aggregated resources, etc. For example, $RO_1$, $RO_2$, $RO_5$, $RO_6$, $RO_{10}$, $RO_{11}$, $RO_{12}$, $RO_{13}$, $RO_{15}$, $RO_{16}$ and $RO_{18}$ (*Figure* 8.1).
- ***balanced***: Based on both the *content and context* of ROs, this solution takes into account the similarity measure defined in Section 7.1, maintaining the appropriate balance between what is the research topic, which researches are involved and how it was built. This solution can be considered more complete and finely tuned than previous ones, but really any of them are good options. For example, $RO_1$, $RO_4$, $RO_6$, $RO_7$, $RO_{11}$, $RO_{15}$ and $RO_{18}$ (*Figure* 8.1).

Moreover, an author may filter ROs according to *publishing date*, *license rights*, *format*, etc. In these cases, the system will obtain a subgraph of *ro-graph* using only the nodes that verify the criteria.

## 8.2 Recommendation 2: Next-Step

**Goal:** Find next research area based on the historical publications of an author

Using all ROs that an author has published, the system is able to predict the topic distribution of the next research. The system create *trending vectors* splitting the topic distribution of each RO by topics. For each *trending vector* a linear regression is calculated to know the probability of that topic in the next research according to the historical information. The system will show to the author a list of ROs related to that distribution of topics to facilitate a first exploration.

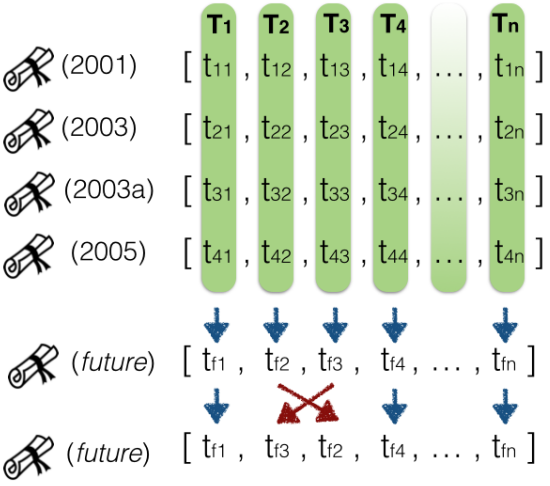


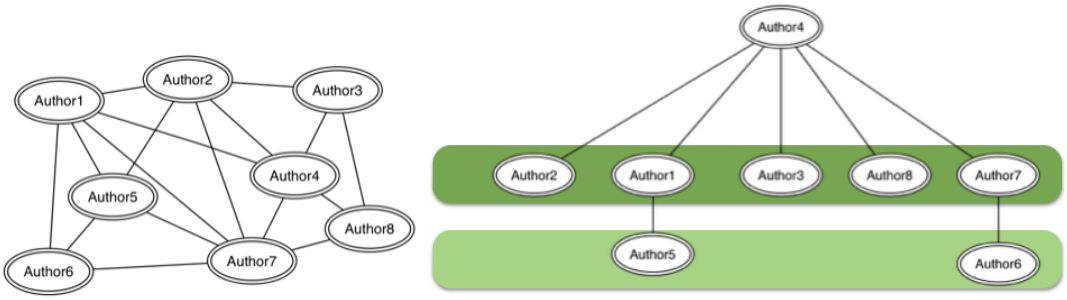

FIGURE 8.2: Trending vectors from publications of an author

Besides, similar to the mutation operator in genetic algorithms, the system introduces a random modification to the distribution of topics to propose new research areas not too far from the current research line showing some statistics about them such as number of ROs published, date of the last one (*hot-topic*), research centers specialized in that field, and so on.

## 8.3 Recommendation 3: Future-Collaborations

**Goal:** Find an interesting list of authors to collaborate based on common research areas

Inspired by the previously mentioned *ro-graph*, the system creates an *author-graph* connecting those authors whose *author-similarity* value is high enough. The similarity between two authors is calculated from their publications, as described in Section 7.2.2,




(a) *authors-graph* (b) *authors-tree*

FIGURE 8.3: Graph and Tree of authors to obtain future collaborations

then it is only based on their research areas identified by their topic distributions and calculated as showed in equation 7.12.

Thus, considering that we have the author-graph shown in *figure* 8.3(a), and that we are interested to know what other authors can be interesting for the author number 4 based on their publications, we'll discover that mainly the authors 2, 1, 3, 8 and 7, and after that the authors 5 and 6 may be interesting according to their research trajectories. Unlike other recommendation systems based on *collaborative-filtering* that use ratings between authors, our system only use the content of publications to make predictions avoiding the previously mentioned *cold-start* problem.

## 8.4 Recommendation 4: Linked-Research

**Goal:** Extract knowledge about researches connecting all the available information

Each of topics discovered in the LDA model describes a research area characterized by its most relevant concepts or words. Moreover, research objects have a topic distribution assigned, so the meta information associated to a research object can be crossed with the previous research areas discovered and other different data sources such as research centers or laboratories where the research was carried out, dates of publication or acceptance, and even public information about authors concerning universities where they have worked or similar. For each author, the system handles digital identifications such as ORCID [55], which provide a persistent digital identifier that distinguishes an author from every other researcher and supports automated linkages to external services such

as Scopus [33], ResearcherID [61] or Linkedin [48], allowing to get information about her/his professional activities.

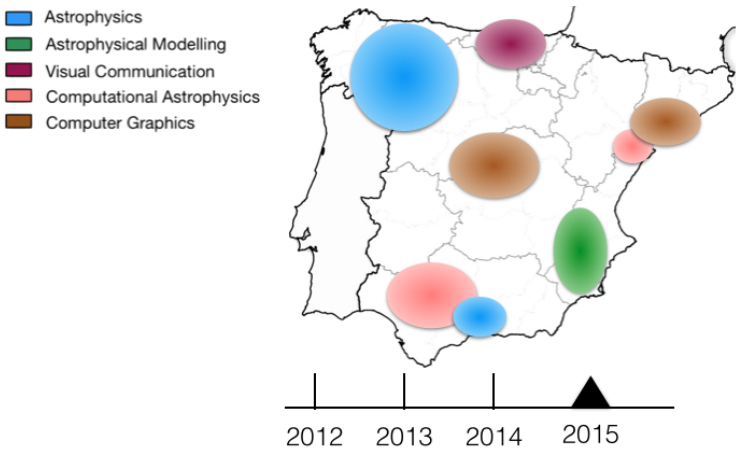


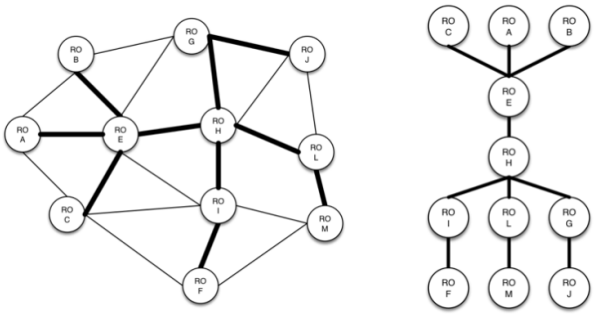

FIGURE 8.4: Simulation of some research topic locations in Spain during 2015

Presenting all this information as *linked data*, i.e. in a connected way maintaining its particular properties and enriching the global information, the system may discover additional knowledge such as the regions in a country where more publications about a specific research area have been done during a period of time, or funds granted for a type of research projects. For instance, in the figure 8.5(b) a map of Spain is showed indicating a potential distribution of some research areas (Astrophysics, Astrophysical Modelling, Visual Communication, Computational Astrophysics and Computer Graphics) in that country based on the research center where the authors work at the time of publication. This could be useful for an author who wants to know how actual is a field or where is the best place to develop or discuss an idea, or who is looking for funds or grants to complete the research that she/he is doing.

The key concept here is to present research objects as *linked* research objects to other different data sources available in the system such as research center location, research grants, author profiles and so on, that can be connected by their meta information. These relations can be interesting to discover hidden information about where, how, who or what are being developed now or in the past.

## 8.5 Recommendation 5: Optimal-Review

**Goal:** Best way to read a group of research objects




(a) *ro-graph* (b) *ro-tree*

FIGURE 8.5: Optimal review of research objects

Suppose that an author has marked a group of researches, i.e. research objects, as interesting for her/his work and she/he needs to read them. As showed in figure 8.5(a), these research objects are A, B, C, E, F, G, H, I, J, L and M. So, she/he can read them in a random order or letting our system define an order based on similarities between them. This situation is similar to the one described in *Route-of-Knowledge*, but now all the research objects have to be used and the list of research objects is not strict. Depending on the similarity measure used to create the graph, the system will offer a *direct*, *consistent* or *balanced* solution as aforementioned.

When the graph is built using only a *content-based* similarity measure, the solution offered will be *direct*. When that measure is only based on *context*, the solution will be *consistent*, and finally, applying a *research object-based* similarity measure, the offered solution will be *balanced*. In any case, the system calculates the *minimum spanning tree* (MST) to connect all the researches. This tree describes an order to read these researches according to the previous criteria.

# Chapter 9

# Conclusions and Future Work

Semantic analysis and its integration in personalization models is one of the most innovative and interesting approaches proposed in the literature to create recommender systems. In this thesis, we have tried to go beyond creating a researcher assistant. The main difference is that we do not only recommend ROs, but also suggest new research lines and routes of knowledge.

The work presented in this thesis is only the base to create a more powerful system able to solve any problem that a researcher can find during her/his research process. The main motivation for future work is the challenge of providing a system with the cultural and linguistic background knowledge that enables interpreting natural language documents, code scripts or even images and reasoning on their content, so as to serve as a basis for the recommendation.

In this line, our similarity measure based on context will take into account not only author profiles, but also style of writing, style of coding, types of aggregated resources and so on. Thus, we will need to particularize how two resources are compared, defining specific methods for images, for scripts, for notes or whatever.

Concerning author-based similarity measure, we will consider different approaches to obtain a common (and highly descriptive) author profile from the set of authors of a research. The centroid of their profiles is one of them, instead of the minimum similarity value between them. Other approach is to define different weights to the authors to obtain a global author-profile measure based on these weights. In this case, we will need to define a criteria to set weights from the meta-information related to a research, e.g.

order of appearance. But may be that criteria is only applied to a publisher, so we will need to consider whether this information (weight or relevance of an author in a research) can be inferred from the context of a research object or explicitly included in the own research object depending on the publisher.

That approach could also be used to build the individual author-profile. Publications wrote only by the author have a higher weight than other wrote by more authors. Usually a research is completed by at least two researchers, so the weight measure should be increased or reduced according to the number of authors. This measure should follow the temporal constraints that we have defined in Section 7.

Following with the improvement of the similarity measures, we think that the similarity measure based on content would be more accurate if the stemming process was more exhaustive. Taking into account that research publications usually follow common templates, our system could detect and learn the patterns that describe these templates to define a list of *stop-words* for each of them. Thus, research publications will not be similar only because they share a common template of publishing.

Including these learning techniques to our system, we could also create a new similarity measure based on the *shape* of a research object, i.e. *template* for textual content or *outline* for image content, to connect research objects with similar structure. This process can be considered as spatial analysis of research objects, but really it is only based on the frequency of terms or pieces of an image, not where the term or the image is located. So, depending on the research method used, e.g. *quantitative, qualitative, applied, correlational or experimental*, two research objects may be more or less similar.

Besides, we will need to express in natural language the meaning of a topic in the sense of topics distributions. In this way, we will offer a richer user experience and the suggestions will be more accurate. Internally, the system can work using terms to describe a topic but externally, i.e. in the author side, a topic should be described by a sentence in natural language easy for authors to be understand.

In any case, future works should only consider as valid the observation or reaction mechanisms based on *implicit feedback*. So, researchers (intended as users of the recommender system) cannot make actions over research objects beyond the authorship. As mentioned in Section 2, this technique does not require any active user involvement, in the sense

that feedback is derived from monitoring and analyzing user's activities, it is only based on the content and features of the resources of the system. In this way, our system will not be limited to relationships created directly by authors, e.g. like/dislike ratings, but also will allow to discover new relationships based on implicit features from both authors and research objects.

# Appendix A

# OAI-DC Metadata XML Schema

```
<schema targetNamespace="http://www.openarchives.org/OAI/2.0/oai_dc/"
        xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
        xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns="http://www.w3.org/2001/XMLSchema"
        elementFormDefault="qualified" attributeFormDefault="unqualified">
<annotation>
  <documentation>
      XML Schema 2002-03-18 by Pete Johnston.
      Adjusted for usage in the OAI-PMH.
      Schema imports the Dublin Core elements from the DCMI schema for unqualified Dublin Core
     .
      2002-12-19 updated to use simpledc20021212.xsd (instead of simpledc20020312.xsd)
  </documentation>
</annotation>
<import namespace="http://purl.org/dc/elements/1.1/"
        schemaLocation="http://dublincore.org/schemas/xmls/simpledc20021212.xsd"/>

<element name="dc" type="oai_dc:oai_dcType"/>
<complexType name="oai_dcType">
  <choice minOccurs="0" maxOccurs="unbounded">
    <element ref="dc:title"/>
    <element ref="dc:creator"/>
    <element ref="dc:subject"/>
    <element ref="dc:description"/>
    <element ref="dc:publisher"/>
    <element ref="dc:contributor"/>
    <element ref="dc:date"/>
    <element ref="dc:type"/>
    <element ref="dc:format"/>
    <element ref="dc:identifier"/>
    <element ref="dc:source"/>
    <element ref="dc:language"/>
    <element ref="dc:relation"/>
    <element ref="dc:coverage"/>
    <element ref="dc:rights"/>
  </choice>
</complexType>
</schema>
```

Listing A.1: XML schema for validating Unqualified Dublin Core metadata associated with the reserved *oai_dc metadataPrefix*

# Appendix B

# OAI-PMH/RSS Harvest Routes

```
/****** ROUTE 1: Slashdot ********/
from("rss:http://rss.slashdot.org/Slashdot/slashdot?" +
 "splitEntries=true&consumer.initialDelay=1000&consumer.delay=2000" +
  "&feedHeader=false&filter=true").marshal().rss().
  setProperty(SOURCE_NAME, constant("slashdot")).
  setProperty(SOURCE_URL, constant("http://rss.slashdot.org/Slashdot/slashdot")).
  to("direct:setCommonRssXpathExpressions").
  to("direct:retrieveByHttpAndSave")

/****** ROUTE 2: UPM ********/
from("oaipmh://oa.upm.es/perl/oai2?initialDelay=1000&delay=60000").
  setProperty(SOURCE_NAME, constant("upm")).
  setProperty(SOURCE_URL, constant("http://oa.upm.es/perl/oai2")).
  to("direct:setCommonOaipmhXpathExpressions").
  setProperty(PUBLICATION_URL,
      xpath("//oai:metadata/oai:dc/dc:relation/text()",String.class).namespaces(ns)).
  to("direct:retrieveByHttpAndSave")

/****** ROUTE 3: UCM ********/
from("oaipmh://eprints.ucm.es/cgi/oai2?initialDelay=1000&delay=60000").
  setProperty(SOURCE_NAME,  constant("ucm")).
  setProperty(SOURCE_URL, constant("http://eprints.ucm.es/cgi/oai2")).
  to("direct:setCommonOaipmhXpathExpressions").
  setProperty(PUBLICATION_URL,
      xpath("//oai:metadata/oai:dc/dc:identifier/text()",String.class).namespaces(ns)).
  to("direct:retrieveByHttpAndSave")

/****** ROUTE 4: SciencePubCo ********/
from("oaipmh://www.sciencepubco.com/index.php/IJAA/oai?initialDelay=1000&delay=60000").
  setProperty(SOURCE_NAME, constant("sciencepubco")).
  setProperty(SOURCE_URL, constant("http://www.sciencepubco.com/index.php/IJAA/oai")).
  to("direct:setCommonOaipmhXpathExpressions").
  setProperty(PUBLICATION_URL,
      xpath("replace(substring-before(concat(string-join(//oai:metadata/oai:dc/dc:relation/
    text(),\";\"),\";\"),\";\"),\"view\",\"download\")",String.class).namespaces(ns)).
  to("direct:retrieveByHttpAndSave")
```

Listing B.1: Example of routes to harvest OAI-PMH and RSS data providers. More details in [11]

# Appendix C

# Research Object Harvester Processor

```
@Component
public class UIAContextGenerator implements Processor {

    private static final Logger LOG = LoggerFactory.getLogger(UIAContextGenerator.class);

    @Override
    public void process(Exchange exchange) throws Exception {

        // UIA Context
        Context context = new Context();

        // Source of data
        Source source = new Source();
        source.setName(exchange.getProperty(AbstractRouteBuilder.SOURCE_NAME, String.class));
        source.setUri(exchange.getProperty(AbstractRouteBuilder.SOURCE_URI, String.class));
        source.setUrl(exchange.getProperty(AbstractRouteBuilder.SOURCE_URL, String.class));
        source.setProtocol(exchange.getProperty(AbstractRouteBuilder.SOURCE_PROTOCOL, String.
    class));
        context.setSource(source);

        // Publication
        Publication publication = new Publication();

        // Metadata
        Reference reference = new Reference();
        reference.setFormat(exchange.getProperty(AbstractRouteBuilder.
    PUBLICATION_METADATA_FORMAT, String.class));
        reference.setUrl("file://"+exchange.getProperty(AbstractRouteBuilder.
    PUBLICATION_REFERENCE_URL, String.class));
        publication.setReference(reference);

        publication.setTitle(exchange.getProperty(AbstractRouteBuilder.PUBLICATION_TITLE,
    String.class));
        publication.setUri(exchange.getProperty(AbstractRouteBuilder.PUBLICATION_URI, String.
    class));
        publication.setFormat(exchange.getProperty(AbstractRouteBuilder.PUBLICATION_FORMAT,
    String.class));
        publication.setLanguage(exchange.getProperty(AbstractRouteBuilder.PUBLICATION_LANGUAGE
    ,String.class));
```

```
33          publication.setPublished(exchange.getProperty(AbstractRouteBuilder.
        PUBLICATION_PUBLISHED, String.class));
34          publication.setRights(exchange.getProperty(AbstractRouteBuilder.PUBLICATION_RIGHTS,
         String.class));
35          publication.setUrl("file://" + exchange.getProperty(AbstractRouteBuilder.
        PUBLICATION_URL_LOCAL, String.class).replace("."+reference.getFormat(), "."+publication.
        getFormat()));
36
37
38          publication.setDescription(exchange.getProperty(AbstractRouteBuilder.
        PUBLICATION_DESCRIPTION, String.class));
39          context.add(publication);
40
41          Iterable<String> iterator = Splitter.on(';').trimResults().omitEmptyStrings().split(
        exchange.getProperty(AbstractRouteBuilder.PUBLICATION_CREATORS, String.class));
42          ArrayList<String> creators = Lists.newArrayList(iterator);
43          publication.setCreators(creators);
44
45
46          Gson gson = new Gson();
47          String json = gson.toJson(context);
48
49
50          exchange.getIn().setBody(json,String.class);
51
52          LOG.debug("Json: {}", json);
53
54      }
55  }
```

Listing C.1: Camel processor that creates research objects from both metadata and related resources . More details in [8]

# Appendix D

# Corpus

| Source | URI | Title | Authors |
|---|---|---|---|
| IJET | *oai:ojs.innovareacademics.in:article/1130* | INSERTION METHOD USING MUSIC NOTES | * Manimuthu,Yamuna |
| IJS | *oai:ojs.innovareacademics.in:article/477* | ARTEMETHER LUMEFANTRINE LOADED LIPOSPHERES EVALUATION OF PROPERTIES OF SOLUTOL HS 15 AND SOLUPLUS ON THE IN VITRO PROPERTIES | * Chime,Salome Amarachi |
| IJSS | *oai:ojs.innovareacademics.in:article/1661* | MUSIC EDUCATION AS A PANACEA FOR NATIONAL DEVELOPMENT | * D.O.A,Ogunrinade |
| IJSS | *oai:ojs.innovareacademics.in:article/1128* | INFLUENCE OF ELECTRONIC MEDIA ON CHILDREN'S PERSONALITY DEVELOPMENT | * Menhas,Rashid<br>* Pir Mehr Ali Shah Arid Agriculture University Rawalpindi,Pakistan |
| IJSS | *oai:ojs.innovareacademics.in:article/1152* | Stakeholder Preference, Dependence and Attitude towards Conservation of Mangrove Eco-System in South-East Coast of India | * CHELLAPPAN,SEKAR<br>* TAMIL NADU AGRICULTURAL UNIVERSITY,COIMBATORE-3 |
| IJBM | *oai:ojs.innovareacademics.in:article/14* | ATTITUDE OF WORKING WOMEN TOWARDS INVESTING IN LIFE INSURANCE WITH SPECIAL REFERENCE TO PRIVATE BANK EMPLOYEES OF COIMBATORE CITY | * S,Vinoth<br>* Associate Professor,RVSIMSR |

| | | | |
|---|---|---|---|
| IJMS | *oai:ojs.innovareacademics.in:article/4003* | SIMILE BETWEEN THE MODUS OPERANDI OF ANALGESIA OF TRAMADOL AND POISON OAK (RHUS TOXICODENDRON) ON FIBROMYALGIA | * Bagchi,Suman<br>* MD (Homoeopathy),SRF(H) at Dr. Anjali Chatterjee Regional Research Institute(H)<br>* Halder,Suman<br>* MD (Homoeopathy),SRF(H) at Dr. Anjali Chatterjee Regional Research Institute(H)<br>* Ex- Assistant Professor of Community Medicine,P.C.M. Homoeopathic Hospital and College<br>* Ghosh,Shubhamoy<br>* MD (Homoeopathy),Head<br>* Roy,Mousumi<br>* BHMS,HMO at Chhurra |
| IJBM | *oai:ojs.innovareacademics.in:article/250* | PERFORMANCES OF INDIAN POSTAL SERVICES | * Anand.,M.B<br>* Asst professor at PESITM,Shivamogga |
| IJSS | *oai:ojs.innovareacademics.in:article/1041* | PUBLIC HEALTH CARE EXPENDITURE IN NIGERIA: CIVILIAN VERSUS MILITARY REGIMES | * Odoh,Vitalis T.<br>* Diamond Bank Plc,Jos<br>* Nduka,Eleanya K.<br>* Department of Economics,University of Nigeria |
| IJAS | *oai:ojs.innovareacademics.in:article/218* | CLINICAL-COMPARATIVE STUDY OF VIRECHAN and PAKSHAGHATARI GUGGULU ON PAKSHAGHAT W.R.S. TO HEMPIPLIGIA. | * kanungo,Neeraj<br>* Gupta,Manoj Kumar<br>* Gaur,Dinesh Singh<br>* Sharma,Shrikrishna<br>* Nigam,U.S<br>* Singh,Vinod kumar |
| IJOE | *oai:ojs.innovareacademics.in:article/2694* | ROLE OF AN IMPORTANCE OF ACTIVITIES IN SCHOOL ENVIRONMENT | * Babu,Anil<br>* Maiwal,Jyoti |
| IJAGS | *oai:ojs.innovareacademics.in:article/1118* | EVALUATION OF MICRO-IRRIGATION, FERTIGATION AND WEED MANAGEMENT IN SUMMER GROUNDNUT | * Mathukia,R. K.<br>* Associate Research Scientist,Department of Agronomy |
| IJET | *oai:ojs.innovareacademics.in:article/1131* | GENERATING DSS GRAPH BY EDGE SUBDIVISION AND EDGE CONTRACTION | * Manimuthu,Yamuna |
| IJHS | *oai:ojs.innovareacademics.in:article/4451* | Following vaccination, Japanese encephalitis (JE) circumstances in Lakhimpur, Assam | * Sharma,Jitendra<br>* District Epidemiologist,Office of the Joint Director of Health Services<br>* Das,J N<br>* District Surveillance officer,Office of the Joint Director of Health Services |

| | | | |
|---|---|---|---|
| IJAGS | *oai:ojs.innovareacademics.in:article/2286* | USE OF AZOLLA BIOFERTILIZER IN POT CULTURE STUDIES WITH PADDY CROP ORYZA SATIVA | * NN,Arumugam<br>* DEPT OF BIOLOGY,GANDHIGRAM RURAL UNIVERSITY GANDHIGRAM-DINDIGUL TAMIL NADU INDIA |
| IJS | *oai:ojs.innovareacademics.in:article/6026* | SIMULTANEOUS ESTIMATION OF IRBESARTAN AND ATORVASTATIN BY Q ABSORPTION RATIO METHOD IN THEIR SYNTHETIC MIXTURE USE IN CARDIAC CONDITION | * Virani,Paras |
| IJHS | *oai:ojs.innovareacademics.in:article/2864* | Relationship between cigarette smoking and body mass index in the Italian population | * De Candia,Gioacchino |
| IJMS | *oai:ojs.innovareacademics.in:article/788* | AN INTRODUCTION OF PROBLEM BASED LEARNING IN IMS, BHU | * Pandey,U.<br>* Mahapatra,T. M. |
| IJET | *oai:ojs.innovareacademics.in:article/981* | THE EFFECT OF SEWAGE CONCENTRATIONS AND MATERIALS OF CONSTRUCTION OF SEWAGE DIGESTER ON BIOGAS PRODUCTION | * Vincent E,Efeovbokhan<br>* Ayodeji A,Ayoola<br>* Omoniyi A,Ayeni<br>* U. Racheal,Essien |
| IJSS | *oai:ojs.innovareacademics.in:article/1132* | LYING HONESTLY FOR GOVERNMENT: LINGUISTIC MANIPULATION AS DISINFORMATION STRATEGY IN NIGERIA | * Agbedo,Christopher Uchenna<br>* Department of Linguistics Igbo and Other Nigerian Languages,University of NIgeria Nsukka Nigeria<br>* Krisagbedo,Ebere Celina |
| IJS | *oai:ojs.innovareacademics.in:article/785* | SYNTHESIS and CHARACTERIZATION OF SERIES LIGANDS AND THEIR COMPLEXES WITH (NI2+) | * Jebur,Miad. Hassan.<br>* Asist . prof .,Chem .Dept |
| IJBM | *oai:ojs.innovareacademics.in:article/4651* | E-Commerce in India - with its whole bag of tricks | * Sundar,Amirtha<br>* National Institute of Technology,Trichy<br>* Vj,Sivakumar<br>* National Institute of Technology,Trichy |
| IJLS | *oai:ojs.innovareacademics.in:article/161* | DIFFERENT MODELS TO EVALUATE ANTIMICROBIAL AGENTS-A REVIEW | * Vashist,Hemraj<br>* Department of Pharmacy,L.R.Institute of Pharmacy<br>* Sharma,Diksha<br>* Department of Pharmacy,L.R.Institute of Pharmacy<br>* Gupta,Avneet<br>* Department of Pharmacy,L.R.Institute of Pharmacy |
| IJSS | *oai:ojs.innovareacademics.in:article/3409* | HUMAN RIGHTS ISSUES IN INDIA - A MAPPING OF DIFFERENT GROUPS | * Duhan,Roshni Dahiya<br>* Maharshi Dayanand University,Rohtak B.P.S.M.University |

| | | | |
|---|---|---|---|
| IJAGS | *oai:ojs.innovareacademics.in:article/80* | ADOPTION STATUS AND FIELD LEVEL PERFORMANCE OF DIFFERENT PROTECTED STRUCTURES FOR VEGETABLE PRODUCTION UNDER CHANGING SCENARIO | * Chatterjee,Ranjit<br>* Uttar Banga Krishi Viswavidyalaya,Pundibari<br>* Mahanta,Sandip<br>* Pal,P.K. |
| IJBM | *oai:ojs.innovareacademics.in:article/245* | CUSTOMER SATISFACTION TOWARDS HOSPITALS , A STUDY ON SELECTED HOSPITALS AT SHIVAMOGGA CITY | * Anand.,M.B<br>* Asst professor at PESITM,Shivamogga<br>* Sudharshan,G.M.<br>* Nagaraja,S.R. |
| IJS | *oai:ojs.innovareacademics.in:article/1150* | EFFECT OF LYOPHILIZATION ON THE PHYSICOCHEMICAL AND PHYSICOTECHNICAL PROPERTIES OF ASPIRIN-LOADED LIPOSPHERES | * Chime,Salome Amarachi<br>* Thaddeus H,Gugu |
| IJET | *oai:ojs.innovareacademics.in:article/630* | OFDM SYSTEMS BASED ON INTER CARRIER INTERFERENCE WITH ASB | * Raman,Ravi<br>* ANNA UNIVERSITY,CHENNAI |
| IJSS | *oai:ojs.innovareacademics.in:article/1620* | Empirical Analysis of Trade Barriers and Economic Growth in Nigeria | * David-Wayas,Onyinye Maria<br>* UNIVERSITY OF NIGERIA,NSUKKA |
| IJAS | *oai:ojs.innovareacademics.in:article/922* | RIVIEW OF SHRINGA , ALABY AND CUPPOING THERAPY | * Katara,Pankaj Kumar<br>* Ch. Brahm Prakash ayurved Charak Sansthan,Khera Dabar Govt Of NCT Delhi |
| IJAGS | *oai:ojs.innovareacademics.in:article/1113* | IRRIGATION AND INTEGRATED NUTRIENT MANAGEMENT IN CASTOR (RICINUS COMMUNIS L.) | * Mathukia,R. K.<br>* Associate Research Scientist,Department of Agronomy |
| IJMS | *oai:ojs.innovareacademics.in:article/154* | PREPARATION OF CHITOSAN STABILIZED OFLOXACIN- GOLD NANO CONJUGATE FOR THE IMPROVED ANTI BACTERIAL ACTIVITY AGAINST HUMAN PATHOGENIC BACTERIA | * Namasivayam,S Karthick Raja<br>* Samrat,K<br>* Ganesh,S |
| IJAS | *oai:ojs.innovareacademics.in:article/533* | A CLINICAL EVALUATION OF GOMUTRA KSHARA COATED SUTRA IN THE MANAGEMENT OF BHAGANDARA (FISTULA- IN -ANO) | * Bhargava,Akhlesh<br>* Yadav,Manoj Kumar<br>* Kushwah,H.K. |
| IJHS | *oai:ojs.innovareacademics.in:article/284* | EFFICACY OF AYURVEDIC DRUGS ON THE 150 PATIENTS OF DIABETIC NEPHROPATHY | * Gupta,Manoj<br>* Lecturer,Department of Roga and Vikriti Vijnana<br>* Chauhan,Ajit Pal Singh<br>* Sharma,Babita<br>* Gaur,Dinesh Singh<br>* Adhikari,S K Das<br>* Singh,Vinod Kumar<br>* Nayak,Subrat Kumar<br>* Urmaliya,Nitin |

| | | | |
|---|---|---|---|
| IJLS | *oai:ojs.innovareacademics.in:article/544* | PHYTOCHEMICAL SCREENING OF MEDICINAL PLANT-MIKANIA CORDIFOLIA AND DETERMINATION OF ITS CHARACTARISTICS. | * Kar,Auditi<br>* Mohammad,Nor<br>* Majumder,Mohammad Sakim<br>* Al-Qayum,Rashed-<br>* Didar Khan,Mohammad<br>* Bhattacharjee,Shovon |
| IJAGS | *oai:ojs.innovareacademics.in:article/746* | THE STATUS OF HIGHLY ALIEN INVASIVE PLANTS IN PAKISTAN AND THEIR IMPACT ON THE ECOSYSTEM: A REVIEW | * Rashid,Mahrine<br>* Haider Abbas,Syed<br>* Rehman,Abdul |
| IJAS | *oai:ojs.innovareacademics.in:article/2672* | HEPATITIS AND PHYSIOLOGY OF LIVER CELLS-A REVIEW | * C,Ugwu Godwin<br>* Chikwendu,Ejere Vincent<br>* Laurete,Okanya Chinagorom<br>* Victor,Egbuji Jude |
| IJAGS | *oai:ojs.innovareacademics.in:article/2561* | COMPATIBILITY OF BEAUVERIA BASSIANA (BALS.) VUILL ISOLATES WITH SELECTED INSECTICIDES AND FUNGICIDES AT AGRICULTURE SPRAY TANK DOSE | * Challa,Murali Mohan<br>* Associate Professor Department of Biotechnology GIT,GITAM University Visakhapatnam |
| IJHS | *oai:ojs.innovareacademics.in:article/3975* | ALL ABOUT YOGA | * Vardhini,R.d.shailima |
| IJS | *oai:ojs.innovareacademics.in:article/347* | SYNTHESIS AND X-RAY CRYSTALLOGRAPHY OF 2,4,6-TRIMETHYL-1,4-DIHYDRO-PYRIDINE-3,5-DICARBOXYLIC ACID DIETHYL ESTER | * Saeed,Sohail<br>* DEPARTMENT OF CHEMISTRY,RESEARCH COMPLEX<br>* Mohamed,Shaaban K. |
| IJAS | *oai:ojs.innovareacademics.in:article/403* | A CLINICAL EVALUATION OF MADHUKADI AND JATYADI TAILA ALONG WITH STANDARD KSHARA-SUTRA THERAPY W.S.R. TO UNIT CUTTING TIME IN THE MANAGEMENT OF BHAGANDARA (FISTULA-IN-ANO) | * Bhargava,Akhlesh |
| IJLS | *oai:ojs.innovareacademics.in:article/4596* | COMBATING MRSA: AN EMPIRICAL STUDY – EVIDENCE FROM PAKISTAN | * Rashid,Anas<br>* Hamdard Institute of Pharmaceutical Sciences (HIPS),Hamdard University Islamabad Campus (HUIC)<br>* Qureshi,Usamah Rashid<br>* Department of Business Studies,Faculty of Economics and Business Studies<br>* Rashid,Aiman<br>* Department of Design and Manufacturing Engineering,School of Mechanical and Manufacturing Engineering (SMME)<br>* Rashid,Hamza<br>* Department of Computer Sciences,Faculty of Natural Sciences |

| | | | |
|---|---|---|---|
| IJOE | *oai:ojs.innovareacademics.in:article/4541* | A Study of teacher's problems of primary schools in rural areas of Mandsaur district | * Mahar,Jaideep |
| IJLS | *oai:ojs.innovareacademics.in:article/539* | CADMIUM CHLORIDE INDUCED CHANGES IN PROTEIN MOLECULES OF THE FRESHWATER FISH CIRRHINUS MRIGALA (HAMILTON) | * K,Veeraiah<br>* K,Jaya Raju<br>* P,Padmavathi<br>* A,Samyuktha Rani<br>* Vivek,Ch |
| IJMS | *oai:ojs.innovareacademics.in:article/534* | RESEARCH ON FORMULATION AND EVALUATION OF INSITU MUCOADHESIVE NASAL GELS OF METOCLOPRAMIDE HYDROCHLORIDE | * Gandhi,Parth Kumar<br>* Department of Pharmaceutics,Kota College of Pharmacy<br>* Rathod,Hemant<br>* Patel,Saurabh<br>* Gandhi,Ronak<br>* Khinchi,M. P.<br>* Agrawal,Dilip<br>* Shrma,Natasha<br>* Kabra,Mahavir |
| IJET | *oai:ojs.innovareacademics.in:article/984* | THE FUTURE AND PROSPECTS OF BIOCHIPS | * Shivakumar,Neeta<br>* S,Poornima<br>* Raghu,Sukanya<br>* Pratibha,. |
| IJS | *oai:ojs.innovareacademics.in:article/4350* | ANTIBACTERIAL ACTIVITY IN DIFFERENT EXTRACTS OF LANTANA CAMARA AGAINST ENTEROPATHOGENS | * Yadav,Shaili<br>* Department of Bioscience and Biotechnology,Banasthali University(304022)<br>* Bhardwaj,Garima<br>* Department of Bioscience and Biotechnology,Banasthali University (304022)<br>* Srivastava,Jyoti<br>* Department of Bioscience and Biotechnology,Banasthali University (304022) |
| IJHS | *oai:ojs.innovareacademics.in:article/398* | EVALUATION OF PHARMACOGNOSTICAL AND PHYTOCHEMICAL PROPERTIES OF THE LEAVES OF PSIDIUM GUAJAVA LINN- BANGALORE VARIETY | * Venkatachalam,Karthikeyan<br>* THE TAMILNADU DR.MGR MEDICAL UNIVERSITY,CHENNAI |
| IJHS | *oai:ojs.innovareacademics.in:article/436* | QUALITY ASSESSMENT PROFILE OF THE LEAVES OF VITIS VINIFERA L. (VITACEAE) – AN IMPORTANT PHYTOTHERAPY COMPONENT OF TROPICAL DISEASES CONTROL | * Venkatachalam,Karthikeyan<br>* THE TAMILNADU DR.MGR MEDICAL UNIVERSITY,CHENNAI |
| IJAGS | *oai:ojs.innovareacademics.in:article/462* | EFFECT OF MOISTURE CONTENT ON SOME PHYSICAL , MECHANICAL PROPERTIES OF LIMA BEAN ( PHASEOLUS LUNATUS ), AN UNDERUTILIZED COMMON FOOD LEGUME | * M,Venkatesh<br>* M,Marimuthu<br>* M,Poongodi |

| | | | |
|---|---|---|---|
| IJAGS | *oai:ojs.innovareacademics.in:article/3469* | EFFECT OF SUPPLEMENTAL IRRIGATION ON WHEAT WATER PRODUCTIVITY UNDER RAINFED ECOLOGY OF POTHOHAR, PAKISTAN | * Asim,Muhammad<br>* PAKISTAN AGRICULTURAL RESEARCH COUNCIL,ISLAMABAD |
| IJLS | *oai:ojs.innovareacademics.in:article/216* | CLINICAL-COMPARATIVE STUDY OF VIRECHAN and PAKSHAGHATARI GUGGULU ON PAKSHAGHAT W.R.S. TO HEMPIPLIGIA. | * Kanoogo,Neeraj<br>* Lecturer,Department of kaya chikitsa |
| IJET | *oai:ojs.innovareacademics.in:article/7092* | Hiding Text within LSB of Image Pixels | * Jain,Rupali<br>* Mittal Institute of Technology,Bhopal |
| IJHS | *oai:ojs.innovareacademics.in:article/985* | PHYTOCHEMICAL EVALUATION OF PLUMBAGO ZEYLANICA: A PREVAILING HERB | * Tyagi,Richa<br>* Menghani,Ekta |
| IJSS | *oai:ojs.innovareacademics.in:article/1134* | 'MY OGA AT THE TOP': PRAGMATIC FAILURES IN THE NIGERIAN INTERLINGUAL COMMUNICATION CONTEXT AND THE LINGUISTIC MECHANISM OF ACCIDENTAL HUMOUR | * Agbedo,Christopher Uchenna<br>* Department of Linguistics Igbo and Other Nigerian Languages,University of NIgeria Nsukka Nigeria<br>* Krisagbedo,Ebere Celina |
| IJBM | *oai:ojs.innovareacademics.in:article/5050* | DETERMINANTS OF EMOTIONAL INTELLIGENCE – THEORETICAL PERSPECTIVE | * E.n,Anju<br>* V,Kubendran |
| IJS | *oai:ojs.innovareacademics.in:article/6028* | DEVELOPMENT AND VALIDATION OF ANALYTICAL METHOD FOR IRBESARTAN AND ATORVASTATIN BY SIMULTANEOUS EQUTION SPECTROSCOPIC METHOD | * Virani,Paras |
| IJAGS | *oai:ojs.innovareacademics.in:article/66* | DEMAND AND SUPPLY PROJECTIONS OF PEARL MILLET IN RAJASTHAN | * Sharma,Shirish<br>* Singh,I.P. |
| IJMS | *oai:ojs.innovareacademics.in:article/566* | THE EFFECT OF METFORMIN 0N CYTOKINES IN IRAQI PATIENTS WITH TYPE 2 DIABETES | * M.K,Yasser<br>* M. R,Abbas<br>* H.M.,Saba |
| IJOE | *oai:ojs.innovareacademics.in:article/6029* | A STUDY OF STUDENT,S PROBLEMS OF PRIMARY SCHOOLS IN RURAL AREAS OF MANDSAUR DISTRICT. | * Mahar,Jaideep |
| IJAS | *oai:ojs.innovareacademics.in:article/3078* | AYURVEDIC MEDICINAL PLANT - SHALA (SHOREA ROBUSTA) (A BIRD'S EYE VIEW) | * Adlakha,Manoj Kumar<br>* Bhargava,Akhlesh Kumar<br>* Kapoor,Ritu<br>* Sharma,L. N.<br>* Singh,Chandan |
| IJAS | *oai:ojs.innovareacademics.in:article/759* | ANCIENT AYURVEDIC THERAPY AGNIKARMA IN CORN (KADAR) | * Bhargava,Akhlesh |
| IJSS | *oai:ojs.innovareacademics.in:article/1195* | POVERTY AND SOCIO-ECONOMIC DEVELOPMENT IN NSUKKA LOCAL GOVERNMENT AREA, ENUGU STATE , SOUTHEASTERN NIGERIA | * Ali,Alphonsus Nwachukwu<br>* University of Nigeria Nsukka,Nigeria.<br>* Ogechi,Agbiogwu |

| | | | |
|---|---|---|---|
| IJHS | *oai:ojs.innovareacademics.in:article/377* | PHARMACOGNOSTICAL, PHYTOCHEMICAL ANALYSIS OF THE LEAVES OF PSIDIUM GUAJAVA LINN– ANAKAPALLI VARIEITY | * Venkatachalam,Karthikeyan<br>* THE TAMILNADU DR.MGR MEDICAL UNIVERSITY,CHENNAI |
| IJLS | *oai:ojs.innovareacademics.in:article/3449* | PREVALENCE OF GASTROINTESTINAL AND RESPIRATORY INFECTIONS IN LAKHIMPUR DISTRICT OF ASSAM, INDIA | * Sharma,Jitendra<br>* District Epidemiologist,Office of the Joint Director of Health Services |
| IJOE | *oai:ojs.innovareacademics.in:article/1017* | RIGHT TO EDUCATION (RTE) Act 2009 | * Pandey,Anamika<br>* Shrivastava,Rachana |
| IJLS | *oai:ojs.innovareacademics.in:article/354* | EFFECT OF CONVOLVULUS PLURICAULIS EXTRACT ON CAFETERIA DIET INDUCED OBESITY IN MICE | * Prasad,Shyam Baboo<br>* Assistant Professor,School of Pharmaceutical Sciences<br>* Sharma,Abhishek<br>* P.G. Scholar,School of Pharmaceutical Sciences<br>* Verma,Hitesh<br>* P.G. Scholar,School of Pharmaceutical Sciences<br>* Yashwant,.<br>* Associate Professor,School of Pharmaceutical Sciences |
| IJAS | *oai:ojs.innovareacademics.in:article/1688* | DESIGN, DEVELOPMENT AND TO FORMULATE ANTIMICROBIAL GEL OF TOONA CILIATA ROEM. LEAVES AND FICUS BENGALENSIS LINN. STEM BARK. | * Singh,Satnam<br>* ASBASJSM College of Pharmacy,Bela |
| IJAS | *oai:ojs.innovareacademics.in:article/460* | HYPOPLASTIC EFFECT OF PROSTOWIN VATI AND VASTIKARMA IN THE MANAGEMENT OF BENIGN PROSTATIC HYPERPLASIA | * Singh,Arun<br>* Bhargava,Akhlesh Kr.<br>* Kushwaha,H.K. |
| IJLS | *oai:ojs.innovareacademics.in:article/410* | FORMULATION OF MAGNETIC NANOPARTICLES AND THEIR APPLICATIONS | * Sailaja,Abbaraju Krishna<br>* Associate Proffesor,RBVRR College of pharmacy |
| IJS | *oai:ojs.innovareacademics.in:article/4923* | COMPATIBILITY AND PROCESSING METHODS STUDY OF FORMULATION OF ARTEMETHER–LUMEFANTRINE FIXED DOSE COMBINATION USING ANALYTICAL TOOLS | * Mustapha,Musibau Aderibigbe<br>* Department of Pharmaceutics and Pharmaceutical Technology. Faculty of Pharmacy. University of Benin. Benin City 300 001,Edo state. Nigeria.<br>* Iwuagwu,Magnus A.<br>* Department of Pharmaceutics and Pharmaceutical Technology. Faculty of Pharmacy. University of Benin. Benin City 300 001,Edo state. Nigeria.<br>* Uhumwangho,Michael U. |

| | | | |
|---|---|---|---|
| IJBM | *oai:ojs.innovareacademics.in:article/940* | ROLE OF MICROFINANCE INSTITUTIONS IN RURAL DEVELOPMENT | * Kumar,N.Prasanna |
| IJAGS | *oai:ojs.innovareacademics.in:article/4366* | Optimization of Irrigation and Fertilizer for Sweet Corn (Zea mays L. var. saccharata Sturt) under Climate Change Conditions | * Mathukia,R. K.<br>* Associate Research Scientist,Department of Agronomy |
| IJOE | *oai:ojs.innovareacademics.in:article/939* | DO LEADERSHIP QUALITIES DETERMINE COMPETENT PRINCIPALS | * Sharma,Sailesh<br>* Institute of Educational Leadership,University of Malaya |
| IJET | *oai:ojs.innovareacademics.in:article/835* | A NOVEL DESIGN APPROACH IN REGULATOR SYSTEM FOR ENERGY CONSERVATION | * K,Yadhari.<br>* Raja.K,Karthik<br>* K,Suresh.<br>* .K,Raja<br>* G,Malathi.<br>* Vanitha.N,Suthanthira |
| IJMS | *oai:ojs.innovareacademics.in:article/338* | EVALUATION OF ANTIDEPRESSANT ACTIVITY OF DIPHENHYDRAMINE IN MICE. | * Taqa,Ghada A. |
| IJOE | *oai:ojs.innovareacademics.in:article/3265* | STUDY OF ROLE OF THE GUARDIANS IN THE MANAGEMENT OF PRIMARY SCHOOLS | * Mahar,Jaideep |
| IJET | *oai:ojs.innovareacademics.in:article/723* | INTERNET MARKETING | * Kumar,N.Prasanna |
| IJOE | *oai:ojs.innovareacademics.in:article/12* | A CASE STUDY OF GIFTED CHILD | * Maharana,Nisha<br>* Principal,Saraswati College of Education |
| IJBM | *oai:ojs.innovareacademics.in:article/725* | GLOBALIZATION AND ITS IMPACT ON INDIAN ECONOMY | * Kumar,N.Prasanna |
| IJOE | *oai:ojs.innovareacademics.in:article/891* | IMPACT OF ACTIVE LEARNING STRATEGIES TO ENHANCE STUDENT PERFORMANCE | * Kumar,Sasi<br>* Centre for research,Tamil University |
| IJBM | *oai:ojs.innovareacademics.in:article/724* | CUSTOMERS' ATTITUDE TOWARDS DEBIT CARDS - A STUDY | * Kumar,N.Prasanna |
| IJOE | *oai:ojs.innovareacademics.in:article/415* | COMMON FIXED POINT THEOREM FOR OCCASIONALLY WEAKLY COMPATIBLE MAPPINGS IN FUZZY METRIC SPACE | * Jauhari,Nitin<br>* Jain,Suman |
| IJBM | *oai:ojs.innovareacademics.in:article/190* | CUSTOMER SATISFACTION TOWARDS HOSPITALS A STUDY ON SELECTED HOSPITALS AT SHIVAMOGGA CITY | * B,Anand M<br>* PESITM,Shivamogga<br>* Sudharshan,G. M.<br>* Nagaraja,S. R. |
| IJMS | *oai:ojs.innovareacademics.in:article/1679* | KNOWLEDGE OF MEDICAL NEGLIGENCE AMONG MEDICAL STUDENTS | * Pandey,U. |
| IJS | *oai:ojs.innovareacademics.in:article/6027* | SIMULTANEOUS ESTIMATION OF IRBESARTAN AND ATORVASTATIN BY FIRST ORDER DERIVATIVE SPECTROSCOPIC METHOD IN THEIR SYNTHETIC MIXTURE USE IN HYPERTENSION CONDITION | * Virani,Paras |
| IJAS | *oai:ojs.innovareacademics.in:article/1136* | AYURVEDA AND MENTAL HEALTH | * Deekshitulu,Balaji |

| | | | |
|---|---|---|---|
| IJLS | *oai:ojs.innovareacademics.in:article/120* | TECHNIQUES USED FOR BIOCHEMICAL INVESTIGATION IN RELATION TO FORENSIC ANALYSIS | * Mehta,Piyush<br>* Dashora,Ashok<br>* Sahu,Deepak<br>* Garg,Rahul Kumar<br>* Agarwal,Piyush<br>* Joshi,Bhavesh<br>* Sharma,Deepak |
| IJMS | *oai:ojs.innovareacademics.in:article/943* | NOVEL MANAGEMENT OF ENDODERMAL SINUS TUMOUR DURING PREGNANCY | * Pandey,U. |
| IJMS | *oai:ojs.innovareacademics.in:article/3353* | PATHOLOGICAL OUTCOME IN THE PATIENTS WITH DIFFERENT AILMENTS: A COMPREHENSIVE STUDY IN LAKHIMPUR DISTRICT, ASSAM | * Sharma,Jitendra<br>* Soni,Monika<br>* Malakar,Mridul<br>* Gupta,Sashi |
| IJMS | *oai:ojs.innovareacademics.in:article/263* | PREVALENCE, ETIOLOGY AND CLINICAL FEATURES OF SKELETAL FLUOROSIS: A CRITICAL REVIEW | * Datta,Pratiti<br>* Saveetha Dental College,Chennai<br>* Datta,Pratyay Pratim<br>* Dept. of Pharmacology,Hi-Tech Medical College and Hospital |
| IJBM | *oai:ojs.innovareacademics.in:article/15* | A STUDY ON FACTOR PATTERNS OF RETAIL STORES INFLUENCING THE BUYING BEHAVIOUR IN COIMBATORE CITY | * Prabu,M. Venkatesh<br>* Assistant Professor,GRG School of Management Studies |
| IJET | *oai:ojs.innovareacademics.in:article/774* | EXTRACTION OF RESIN FROM AGRO-INDUSTRIAL WASTES | * Priya,Gomathi |
| IJHS | *oai:ojs.innovareacademics.in:article/2167* | Stress Reduction Through Listening Indian Classical Music | * P V,BALAJI DEEKSHITULU |
| IJHS | *oai:ojs.innovareacademics.in:article/1579* | VEDIC LIFE STYLE IN STRESS CONTROL | * P V,Dr.BALAJI DEEKSHITULU |
| IJOE | *oai:ojs.innovareacademics.in:article/13* | A STUDY OF EMOTIONAL INTELLIGENCE OF HIGHER SECONDARY SCHOOL TEACHERS OF MADHYA PRADESH | * Maharana,Nisha<br>* Principal,Saraswati College of Education |
| IJS | *oai:ojs.innovareacademics.in:article/919* | PHYTOCHEMICAL COMPOSITIONS AND ORGAN WEIGHT EFFECTS OF MUCUNA SLOANEI (FABACEAE) IN ALBINO RATS (RATTUS NOVERGICUS) | * C,Ugwu Godwin<br>* C,Ejere Vincent.<br>* C,Ejere Vincent.<br>* L,Okanya Chinagorom.<br>* Jude,Egbuji<br>* L,Okanya Chinagorom.<br>* Jude,Egbuji |
| IJLS | *oai:ojs.innovareacademics.in:article/4936* | An overall review on Obesity and its related disorders | * Sailaja,Abbaraju Krishna<br>* Associate Proffesor,RBVRR College of pharmacy |
| IJET | *oai:ojs.innovareacademics.in:article/789* | A NOVEL ANALYSIS ON LIGHT ESCAPING FROM BLACK HOLE | * Kaliyannan,Raja<br>* S,Karthick.<br>* .K,Raja<br>* K,Yadhari.<br>* S,Kalpanadevi<br>* Vanitha.N,Suthanthira |

| | | | |
|---|---|---|---|
| IJSS | *oai:ojs.innovareacademics.in:article/1185* | PRELIMINARY STUDY ON THE USE OF SOUND AND ACOUSTICS IN IGBO CULTURAL COMMUNICATION | * Ahamefula,Ndubuisi Ogbonna<br>* Department of Linguistics,Igbo and Other Nigerian Languages<br>* Okoye,Chinenye L.<br>* Onwuegbuchunam,Marcellus O.<br>* Uzoigwe,Benita C.<br>* Nneji,Ogechukwu M. |

TABLE D.1: List of Research Objects used during evaluations

# Bibliography


[1] Agarwal, D. and Chen, B.-C. (2009). Regression-based latent factor models. In *Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining (2009)*, page 19.

[2] Ahmed, A., Aly, M., Gonzalez, J., Narayanamurthy, S., and Smola, A. J. (2012). Scalable inference in latent variable models. In *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*, page 123.

[3] Allard, D., Comunian, a., and Renard, P. (2012). Probability Aggregation Methods in Geoscience. *Mathematical Geosciences*, 44(5):545–581.

[4] Apache (2015a). Apache Camel. *http://camel.apache.org*.

[5] Apache (2015b). Apache Spark. *https://spark.apache.org/docs/latest/mllib-clustering.html#latent-dirichlet-allocation-lda*.

[6] Aseervatham, S. and Bennani, Y. (2009). Semi-structured document categorization with a semantic kernel. *Pattern Recognition*, 42(9):2067–2076.

[7] Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. (2009). On Smoothing and Inference for Topic Models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 27–34.

[8] Badenes, C., Gonzalez, R., and Corcho, O. (2015a). Harvester Application. *https://github.com/cabadol/epnoi-harvester*.

[9] Badenes, C., Gonzalez, R., and Corcho, O. (2015b). OAI-PMH Analyzer. *https://github.com/cabadol/oaipmh-analyzer*.

[10] Badenes, C., Gonzalez, R., and Corcho, O. (2015c). OAI-PMH Camel Component. *https://github.com/cabadol/camel-oaipmh*.

[11] Badenes, C., Gonzalez, R., and Corcho, O. (2015d). OAI-PMH/RSS Hoarder Client. *https://github.com/cabadol/epnoi-hoarder*.

[12] Belhajjame, K., Zhao, J., Garijo, D., Hettne, K., Palma, R., Corcho, O., Gómez-Pérez, J.-M., Bechhofer, S., Klyne, G., and Goble, C. (2014). The Research Object Suite of Ontologies: Sharing and Exchanging Research Data and Methods on the Open Web. *arXiv preprint arXiv: 1401.4307*, (February 2014):20.

[13] Bergmann, F. T., Adams, R., Moodie, S., Cooper, J., Glont, M., Golebiewski, M., Hucka, M., Laibe, C., Miller, A. K., Nickerson, D. P., Olivier, B. G., Rodriguez, N., Sauro, H. M., Scharm, M., Soiland-Reyes, S., Waltemath, D., Yvon, F., and Le Novère, N. (2014). COMBINE archive and OMEX format: one file to share all information to reproduce a modeling project. *BMC Bioinformatics*, 15(1):369.

[14] Betz, F. (2011). Managing Science. *Knowledge Management*, page 190.

[15] Blei, D., Carin, L., and Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6):55–65.

[16] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022.

[17] Breese, J. S., Heckerman, D., and Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th conference on Uncertainty in Artificial Intelligence*, pages 43–52.

[18] Burke, R. (2007). Hybrid web recommender systems. *The adaptive web*, pages 377–408.

[19] Cai, D., Mei, Q., and Han, J. (2008). Modeling Hidden Topics on Document Manifold Categories and Subject Descriptors. In *In Proceedings of the ACM conference on Information and knowledge management*, pages 911 – 920.

[20] Celikyilmaz, a., Hakkani-Tur, D., and Tur, G. (2010). LDA Based Similarity Modeling for Question Answering. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*, pages 1–9.

[21] Codehaus (2015). Groovy Language. *http://www.groovy-lang.org*.

[22] Dagan, I., Lee, L., and Pereira, F. C. N. (1999). Similarity-Based Models of Word Cooccurrence Probabilities. *Machine Learning*, 34(1-3):43–69.

[23] DCMI (2015a). DCMI Metadata Terms. *http://dublincore.org/documents/dcmi-terms/*.

[24] DCMI (2015b). Dublin Core Metadata Initiative. *http://dublincore.org*.

[25] Deb, K. and Agrawal, R. B. (1994). Simulated Binary Crossover for Continuous Search Space. *Complex Systems*, 9:1–34.

[26] Deb, K. and Goyal, M. (1996). A Combined Genetic Adaptive Search (GeneAS) for Engineering Design. *Computer Science and Informatics*, 26(1):30–45.

[27] Deb, K. and Jain, H. (2013). An Evolutionary Many-Objective Optimization Algorithm Using Reference-point Based Non-dominated Sorting Approach, Part I: Solving Problems with Box Constraints. *Ieeexplore.Ieee.Org*, 18(c):1–1.

[28] Deb, K. and Tiwari, S. (2008). Omni-optimizer: A generic evolutionary algorithm for single and multi-objective optimization. *European Journal of Operational Research*, 185(3):1062–1087.

[29] Delgado, J. and Ishii, N. (1999). Memory-based weighted-majority prediction. *SIGIR Workshop Recomm. Syst. Citeseer*.

[30] Deshpande, M. and Karypis, G. (2004). Recommendation Algorithms. *ACM Transactions on Information Systems*, 22(1):143–177.

[31] Dijkstra, E. W. (1959). A note on two problems in connection with graphs. *Numerische Mathematik*, 1(1):269–271.

[32] EIP (2015). Enterprise Integration Patterns. *http://www.enterpriseintegrationpatterns.com*.

[33] Elsevier (2015). Scopus. *http://www.scopus.com*.

[34] Getoor, L. and Sahami, M. (1999). Using probabilistic relational models for collaborative filtering. *Workshop on Web Usage Analysis and User Profiling*.

[35] Goldberg, D., Nichols, D., Oki, B. M., and Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70.

[36] Gunawardana, A. and Meek, C. (2009). A Unified Approach to Building Hybrid Recommender Systems. In *Proceedings of the Third ACM Conference on Recommender Systems*, pages 117–124.

[37] Hasan, S., O'Riain, S., and Curry, E. (2012). Approximate Semantic Matching of Heterogeneous Events. *6th ACM International Conference on Distributed Event-Based Systems (DEBS 2012)*, pages 252–263.

[38] Hettne, K. M., Dharuri, H., Zhao, J., Wolstencroft, K., Belhajjame, K., Soiland-Reyes, S., Mina, E., Thompson, M., Cruickshank, D., Verdes-Montenegro, L., Garrido, J., de Roure, D., Corcho, O., Klyne, G., van Schouwen, R., Hoen, P. a. C. T., Bechhofer, S., Goble, C., and Roos, M. (2013). Structuring research methods and data with the Research Object model: genomics workflows as a case study. *Biomedical Semantics*, page 39.

[39] Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1-2):177–196.

[40] Innovare (2015). Innovare Academic Sciences. *http://innovareacademics.in*.

[41] Investigacion, K. (2015). JMetal. *http://jmetal.sourceforge.net*.

[42] Kim, Y. and Shim, K. (2011). TWITOBI: A recommendation system for Twitter using probabilistic modeling. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 340–349.

[43] Kim, Y. and Shim, K. (2014). TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation. *Information Systems*, 42:59–77.

[44] Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8):42–49.

[45] Lagoze, C. and Sompel, H. V. D. (2003). The open archives initiative protocol for metadata harvesting. *Library hi tech*, 21(2):118–128.

[46] Le Novère, N., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., Crampin, E. J., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J. L., Spence, H. D., and Wanner, B. L. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature biotechnology*, 23(12):1509–1515.

[47] Liagkouras, K. and Metaxiotis, K. (2013). An Elitist Polynomial Mutation Operator for improved performance of MOEAs in Computer Networks. *Computer Communications and Networks (ICCCN), 2013 22nd International Conference on*, pages 1–5.

[48] Linkedin (2015). Linkedin. *https://www.linkedin.com*.

[49] Mesirov, J. P. (2010). Computer Science: Accesible Reproducible Research. *Science*, 327(5964):1–5.

[50] Mild, A. and Reutterer, T. (2001). Collaborative Filtering Methods for Binary. In *Proceedings of the 6th International Computer Science Conference on Active Media Technology*, pages 302–313.

[51] MuleSoft (2015). Mule ESB. *https://www.mulesoft.com/platform/soa/mule-esb-open-source-esb*.

[52] OAI-PMH (2015a). OAI Data Provider Registry. *http://www.openarchives.org/pmh/registry/ListFriends*.

[53] OAI-PMH (2015b). Open Archive Initiative Protocol for Metadata Harvesting. *https://www.openarchives.org/OAI/openarchivesprotocol.html*.

[OpenArchive] OpenArchive. OAI-PMH XML Schema Definition. *http://www.openarchives.org/OAI/openarchivesprotocol.html#OAIPMHschema*.

[55] Orcid (2015). ORCID. *http://orcid.org*.

[56] Osborne, F., Scavo, G., and Motta, E. (2014). Identifying Diachronic Topic-Based Research Communities by Clustering Shared Research Trajectories. *Lecture Notes in Computer Science*, 8465:114–129.

[57] Patra, B. K., Launonen, R., Ollikainen, V., and Nandi, S. (2015). A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data. *Knowledge-Based Systems*, 82:163–177.

[58] Pavlov, D. and Pennock, D. (2002). A maximum entropy approach to collaborative filtering in dynamic, sparse, high-dimensional domains. *In Proceedings of Neural Information Processing Systems*, pages 1441–1448.

[59] Ranjan, R. and Gneiting, T. (2008). Combining probability forecasts. *International Journal of Forecasting*, 27(2):208–223.

[60] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). GroupLens : An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186.

[61] Reuters, T. (2015). ResearcherID. *http://www.researcherid.com*.

[62] Rus, V., Niraula, N., and Banjade, R. (2013). Similarity Measures Based on Latent Dirichlet Allocation. In *Computational Linguistics and Intelligent Text Processing*, pages 459–470. Springer US.

[63] Sahu, P. K. (2013). *Research Methodology: A Guide for Researchers In Agricultural Science, Social Science and Other Related Fields*. Springer US.

[64] Salakhutdinov, R. and Mnih, A. (2007). Probabilistic Matrix Factorization. In *Neural Information Processing Systems Conference*, pages 1–8.

[65] Satopää, V. a., Baron, J., Foster, D. P., Mellers, B. a., Tetlock, P. E., and Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2):344–356.

[66] Singh, A. and Singh, B. (2014). Procedure of Research Methodology in Research Studies. *European International Journal of Science and Technology*, 3(9):79–85.

[67] Spring (2015). Spring Integration. *http://projects.spring.io/spring-integration/*.

[68] Sutherland, W. J., Mitchell, R., and Prior, S. V. (2012). The role of 'conservation evidence' in improving conservation management. *Conservation Evidence*, 9(11):1–2.

[69] Tintarev, N. and Masthoff, J. (2011). *Recommender Systems Handbook*, volume 54. Springer US.

[70] W3C (2015). XML Path Language. *http://www.w3.org/TR/xpath/*.

[71] Waltemath, D., Adams, R., Beard, D. a., Bergmann, F. T., Bhalla, U. S., Britten, R., Chelliah, V., Cooling, M. T., Cooper, J., Crampin, E. J., Garny, A., Hoops, S., Hucka, M., Hunter, P., Klipp, E., Laibe, C., Miller, A. K., Moraru, I., Nickerson, D., Nielsen, P., Nikolski, M., Sahle, S., Sauro, H. M., Schmidt, H., Snoep, J. L., Tolle, D., Wolkenhauer, O., and le Novère, N. (2011). Minimum information about a simulation experiment (MIASE). *PLoS Computational Biology*, 7(4):5–8.

[72] Wang, C. and Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 448–456.

[73] Wang, J., Jin, B., and Li, J. (2004). An Ontology-Based Publish / Subscribe System. In *In Middleware*, pages 232–253. Springer US.

[74] Wf4ever (2014). Research Object Bundle. *https://researchobject.github.io/specifications/bundle/*.

[75] Yildirim, H., Yildirim, H., Krishnamoorthy, M. S., and Krishnamoorthy, M. S. (2008). A random walk method for alleviating the sparsity problem in collaborative filtering. In *Proceedings of the 2008 ACM conference on Recommender systems - RecSys '08*, pages 131–138.

[76] Zhou, A., Qu, B.-Y., Li, H., Zhao, S.-Z., Suganthan, P. N., and Zhang, Q. (2011a). Multiobjective evolutionary algorithms: A survey of the state of the art. *Swarm and Evolutionary Computation*, 1(1):32–49.

[77] Zhou, K., Yang, S.-h., and Zha, H. (2011b). Functional Matrix Factorizations for Cold-Start Recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 315–324.