International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 6 Issue 7|| July 2017 || PP. 72-84

An Assisted Literature Review using Machine Learning Models to Identify and Build a Literature Corpus

Ronald Brisebois[1], Alain Abran[1], Apollinaire Nadembega[2], Philippe N'techobo[3]
[1]*École technologie supérieure, Université du Québec, Canada*
[2]*Network Research Lab., University of Montreal, Canada*
[3]*École Polytechnique de Montréeal, Canada*
Corresponding Author: Ronald Brisebois

Abstract: *With the evolving and interdisciplinary nature of research, there is a need to facilitate and assist researchers in the manual process of building a literature review. This paper proposes an assisted literature review prototype based on machine learning models (MLM) to discover, rank and recommend the relevant papers. Using text and data mining models, MLM and a classification model, all of which learn from researchers' annotated data and semantic enriched metadata, assisted literature review helps researchers identify, rank and select papers. This prototype evaluates papers and bibliographic attributes in order to determine their relevancy and aggregates all relevant contents into an assisted literature review object. This paper presents the MLM and related algorithms that:*
1. Identify the relevant papers harvested from the web and other sources for building the Literature Corpus;
2. Obtain the Literature Corpus radius by calculating the distance of each paper to the center of the Literature Corpus defined for a specific topic, concept or area of research.
The performance, in terms of accuracy, was evaluated and compared to other approaches using a number of assisted literature review prototype simulations with a manual literature review metadata as input parameters.
Keywords**:** *assisted literature review, literature review, machine learning, semantic topic detection, text and data mining*

Date of Submission: 18-07-2017 Date of acceptance: 05-08-2017

I. Introduction

The huge volume of scientific publications available is becoming an issue for researchers [1]: given that their time is limited, it is becoming impossible for researchers to read and carefully evaluate every publication within their own specialized field.

A manual literature review (LR) process is very labor intensive, and the time that researchers must dedicate to searching for literature will vary according to their research topic. For instance, it is estimated that a decent LR for a dissertation takes three to six months to complete. In their academic process, postgraduate students in all disciplines need to be able to write an accurate LR. Whether a short review as an assignment in a Master's program, or a full-length LR for a PhD thesis, students find it difficult to produce a LR with all of the relevant papers. Researchers also have to stay aware of newly published papers on related topics to produce a meaningful LR. According to [2,3], an LR process consists in locating, appraising and synthesizing the best available empirical evidence to answer specific research questions. An LR will look at as much existing research as is feasible and will review scholarly papers and theses in the relevant area.

To manually find sources of content for the LR, the first step is to identify the relevant topics or concepts and prioritize them. A way to identify the relevant ones is to analyze the lists of references to see which are frequently cited and how often. This requires ranking the LR references.

With the massive increase in digital content and widespread use of search engines, the number of returned results can be tremendous—which then makes it challenging to select only the papers relevant to the LR topic. In the context of scientific content, the ranking algorithms for content evaluation are referred to as scientometrics or bibliometrics [4-18].

Semantic metadata allow more accurate searching than keywords and may help to get better relevant results for an assisted literature review (ALR). Semantic metadata can be extracted using text and data mining (TDM) algorithms. TDM, machine learning models (MLM) have been designed to learn from papers and researchers' annotated papers and to identify relevant papers for a specific topic and research field.

This paper, define and analyse an assisted literature review prototype designed to reduce reading load by pointing the researcher to a recommended selection of papers. This paper proposes an ALR prototype, i.e., a set of TDM and MLM algorithms for searching, discovering, ranking and recommending papers for the researcher.

II. Related Works

This section presents the related works in the following sequence:

1. Ranking of scientific papers
2. Text and data mining, and more specifically:

a. Machine learning models (MLM)
b. Automatic multi-documents summarization for ALRs

3. Assisted literature review object (ALRO)

2.1 Ranking of scientific papers

The proliferation of scientific publications and the online availability of repositories make it challenging for researchers to produce and maintain a LR for specific research fields. Two means of quantitatively evaluating scientific research output are discussed in the literature: peer-review and citation-based bibliometrics indicators. The main limitation of peer-review-based approaches is the subjectivity of evaluators, while citations-based approaches have been criticized for having a scope limited to academia and neglecting the broader societal impact of research [13].

According to the literature, citation analysis is widely used to measure scientific papers and their impact. Recently some iterative processes, such as PageRank, have been applied to citation networks to perform this function. The PageRank algorithm also has some limitations: for example, recent papers not yet cited do not appear in the top level of results. Furthermore, the links between papers are oriented to a single direction: from a citing paper to cited papers.

Scientific paper ranking should also depend on the venue, the location of publication, the year, the author and the citation index. Some works in the field of scientific impact evaluation [4,17,15,8] address the ranking of universities and research teams.

For this research, many existing approaches for scientific paper ranking have been evaluated [3,5,6,10,12,13,15,17,18]. They suffer from a number of limitations:

1. Most existing approaches ignore the papers index.
2. Most approaches only use the citations count.
3. Most approaches do not take into account the Social Level Metric, the category or polarity of citations.

2.2 Text and data mining

In scientific research, documents (such as journal papers, conference proceedings or research reports) have a specific organization and relevant sections that are different from other types of documents such as narrative text [19]. The purpose of a text summarizer is to select the most important facts and present them in a sensible order while avoiding repetition [20]. However, scientific papers frequently contain repeated expressions and sentences. Consequently, narrative text summarization approaches are not adequate for summarizing scientific papers for an ALR.

2.2.1 Machine learning models

MLM is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. MLM explores the definition and study of algorithms that can learn from and make predictions on data. According to literature, the field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.

There are three different axes for MLM:

1. Text and data mining: using historical data to improve decisions
2. Software algorithms that are difficult to program by hand
3. User modeling of assistants

In the context of TDM, MLM is used mainly for metadata enrichment and literature review refinement in the context of ALR.

For example, Carlos and Thiago [2] developed a supervised MLM-based solution for text mining scientific articles using the R language in "Knowledge Extraction and Machine Learning" based on social network analysis, topic models and bipartite graph approaches. Indeed, they defined a bipartite graph between documents and topics that makes use of the Latent Dirichlet Allocation topic model.

In regards to the classification and ranking problem, there are different MLM. To determine which model performs best, the best way remains the use of prototypes.

An MLM can also be dynamic, meaning that it can train itself on the analysis of new data. In the case of MLM's K-means clustering algorithm, the data would be classified into clusters and any new metadata and data would clarify the cluster boundaries, thus improving the model's ability to classify accurately.

2.2.2 Automatic multi-document summarization for assisted literature review

Several approaches have been proposed for scientific paper summarization [21-26,2,27-30]. For an ALR, numerous publications need to be analyzed and summarized: this is referred to as multi-document summarization. In the context of scientific research, given a set of scientific papers, multi-document summarization can be used to generate an ALR. According to [31], there are two main styles:

1. A descriptive LR presents a critical summary of a research domain: it summarizes individual papers/studies and provides more information about each one, such as its research methods and results.
2. An integrative LR focuses on the ideas and results extracted from a number of research papers and provides fewer details about individual papers/studies.

2.3 Assisted literature review object

We have coined the term "assisted literature review object" (ALRO) to refer to a component type that includes many types of metadata and content related to the researchers' specific requests; for example, an ALRO may enrich an ALR with a video or speech that facilitates understanding of the topic of a paper. Indeed, an ALRO is built for a given research topic and differs according to the selection parameters, paper annotations and the time of the request. In other words, it is dynamic, and it aggregates data and enriches metadata about a given ALR to help researchers learn about their field more quickly. Very few works have examined ALRO as defined in this way. In one of these works, Dunn et al. [21] present the results of their effort to integrate statistics, text analytics and visualization in a prototype interface for researchers and analysts. Their prototype system, called Action Science Explorer (ASE), provides an environment for demonstrating principles of coordination and conducting iterative usability tests with interested and knowledgeable researchers. According to these authors, ASE is designed to support exploration of a collection of papers by rapidly providing a summary, while identifying key papers, topics and research groups. The first drawback of ASE is that it does not propose an algorithm or model for evaluating a scientific paper's relevancy to its research field, but uses only the paper's bibliometric ranking.

The main drawbacks of existing approaches are as follows:

1. Scientific research papers have a specific structural organization different from that of other types of documents such as narrative or biographical texts.
2. Most of the existing approaches focus only on single paper summarization.
3. Existing works underutilized the annotations, research domain, specific topic, matching keywords and subject of research.

In this research work, we address several limitations of existing approaches [21,31,30,32-36] for the design of a better literature review for researchers.

III. Assisted Literature Model

This section first presents an overview of the prototype model. The various MLM designed for the prototype will then be described.

3.1 Workflows of manual and assisted literature reviews

The workflow of a manual LR is presented in Fig. 1. Within these figures, the white boxes represent manual activities while the shaded ones represent automated activities.

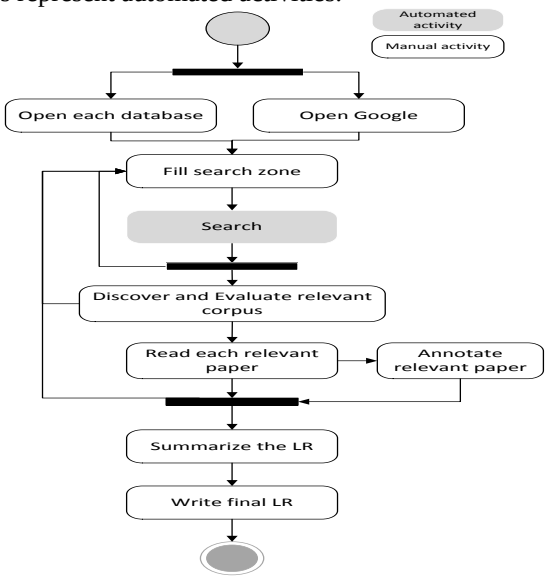

Figure 1: Workflow of a manual LR

3.2 Overview of the prototype of an assisted LR

A literature search has to be systematic and evaluative: it should assess each paper to determine its ranking and whether or not it is worth including in the LR. One of the aims of an ALR is to reduce the reading load by enabling the researcher to exploit only a relevant papers. This prototype uses as inputs:

1. A universal research document repository (URDR)
2. The papers annotated by the researcher and previous researchers.

3.3 SEARCH & REFINE ALR

The Search & Refine ALR consists of seven steps.

1. Identify, Refine & Notify ALR's Selection

This first step identifies and refines, in an interactive process, researcher selection (RS) metadata in order to provide an ALR that meets researcher requirements; it also notifies the researcher when new paper matching with its RS metadata is published.

A secondary objective of this step is to formulate the research questions. The metadata used to identify an RS are defined in two sections:

a. Document Common Metadata section (top part of Table 1)
b. Researcher Annotations section (bottom part of Table 1)

The researcher can iterate this first step as necessary to complete the ALR or when there is a new paper to be added.

Table 1: Researcher selection (RS) metadata

Number	Metadata	Description
A. Document Common Metadata		
1	Discipline	Selection of the discipline related to the ALR
2	Main Topic	The main topic is one of the most important metadata for building the ALR. It should be as specific as possible.
3	Literature Corpus Radius	The Literature Corpus Radius (LCR) is used to build other algorithms; it is the main concept that makes it possible to refine the selection of research documents to be included in the ALR.
4	Keywords	The researcher has to identify keywords representative of the ALR.
5	Harvesting Date	Date of document harvesting
6	Creation Date	Date of document creation
7	Title	Title of the ALR
8	MLTC - Mix Literature Temporal Coverage	The MLTC is very crucial to building and refining the ALR. It has two indicators: 1 - Number of years covered by the search 2 - Percentage of documents outside this time range to be included.
9	Description	A brief description of the research project such as a paper abstract
10	Languages	The researcher has to choose the language of the documents to be included in the corpus of interest.
11	# of References	The number of references that the ALR should consider.
B. Researcher Annotations Metadata		
12	Key Findings	The Key Findings are annotations regarding important findings in the document identified by the researcher.
13	Free Tags	The researcher may place tags on a document in order to remember some information about it.
14	Personal Notes	The researcher may attach notes to a document in order to remember some information about it.
15	Pre-defined Tags	These are predefined metadata to help the researcher to track the status of the relevant document.

2. Discover Relevant Literature & Manage Personal Metadata

From the growing cluster of papers, a literature corpus that meets the RS metadata is identified. Any papers tagged by the researcher as "Relevant for the ALR" will be included. The paper relevancy is measured thanks to dynamic topic based index (DTb index) that is computed making used of TDM and MLM approaches.

3. Evaluate, Organize & Index the Relevant Literature

A subset of relevant papers is created in order to define the ALR Corpus based on the literature corpus radius index (LCR index). In contrast to Literature Corpus which denotes all the papers of a specific research topic, the ALR Corpus denotes only the papers of a Literature Corpus which meets RS metadata for an ALR.

4. Enrich & Summarize the Literature Review

The ALRO is produced through text summarization and subject extraction.

5. Synthesize & Clusterize the ALR Structure & Citations

All the relevant documents are synthesized and organized into clusters related to the LCR index.

6. Generate & Visualize the ALR

In this step, the recommended papers in the Literature Corpus are generated and visualized. Assisted generation of the recommended papers helps the researcher examine the coherence of the ALR and iterate the ALR process. At any moment, the researcher can add to the relevant papers list that will be part of the final ALR.

7. Metadata-based Literature & Research Alerts

New relevant papers or new metadata related to the ALR are detected in this last step.

3.4 ASSIST & RECOMMEND ALR

Assist & recommend ALR allows refining the ALR through two sets of steps (S1 and S2). Numbers 1 to 5 in the bottom-right corner of many of the boxes denote the MLM designed to identify a specific corpus, evaluate document relevancy or define learning models that are required by the prototype for obtaining the ALR objects.
The following sources are used to build the suggested list of ALR papers:

1. The list of papers generated by the step 'ALR Refine & Recommendation - MLM 4' according to the RS. This list includes the LCR threshold indicated by the gray circle (papers in blue).
2. The annotated papers from the researcher (RAs) – papers in red.
3. The papers identified by the Mix Literature Temporal Coverage (MLTC) from the RS – papers in yellow.
4. The universal research document repository (URDR).

Each corpus in Fig. 2 is shown as a circle whose horizontal axis represents the LCR line. Note that the origin of this axis is not explicitly visible. Indeed, the center of each circle denotes the origin of the horizontal axis going off toward the right or left, but the center is hidden by the type of metadata (RS or RA) used to select the corpus. What is more important is to position a paper at the correct distance from the center according to its LCR index. The LCR index of a paper is defined as the similarity between the RS metadata and that paper's metadata such as title, topics, abstract and keywords. It measures the semantic relevancy of a paper according to the RS.

The Literature Corpus contains all the papers regardless of their LCR index and the type of selection metadata (i.e., RSs or RAs). The papers within corpus radius are those located at the surface (forming a disc) of a circle with the specific corpus radius. We refer to the radius of this specific circle as the Corpus Radius (see Fig.2). The Corpus Radius may be defined as the delimiter of the Literature Corpus suggested to the researcher for the ALR on the basis of the researcher's selections and annotations. The goal is to limit the number of papers to those that are relevant. The RA selection criteria consist of notes, tags and key findings mentioned by the researcher.

To illustrate, consider the papers in the corpus radius called "**Papers relevant to ALR**" (disk with blue rectangles at the top of Fig. 2): all the papers within the gray disc are URDR papers whose LCR index is less than or equal to 2; the LCR threshold is set at 2.

3.5 Assisted Literature Review Object (ALRO)

The concept of the assisted literature review object (ALRO) is useful for managing ALRs. It is basically a component type that includes many types of information related to the LR. The Entity Matrix has been modified with the addition of a new component type: ALRO [37]. An ALRO aggregates all objects and relationships related to the creation of an ALR. It can be shared by researchers or re-used to accelerate research findings.
The prototype proposes different indexes to evaluate the relevance and importance of am ALRO for a specific researcher; for example, the DTb index takes into account:

1. Topic-based, Text-based and Reference-based approaches
2. Author-level, Co-author-level, Venue-level, Social-level, Affiliation-level metrics.

Several supervised MLM-based metadata extraction methods are available for automatic integration of metadata into bibliographic manager tools such as Endnote. In this work, which takes a rules-based approach, a supervised MLM is used [3].
Additional metadata about authors and researchers need to be identified or computed. Author metadata is usually the basis of a search for document relevancy detection. They help to gain insights about author' publications.

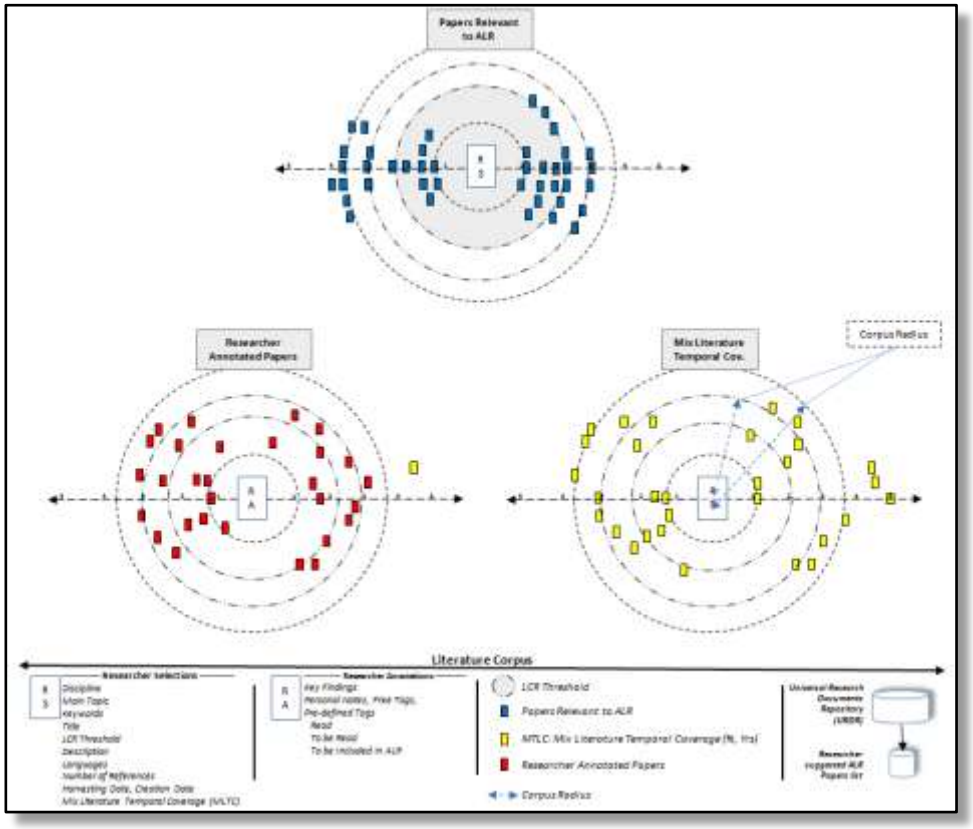

Figure 2: Sources used to build the suggested list of ALR papers

IV. The Prototype Processes Description

This section presents the MLM of the prototype. For an improved understanding of Steps 1 and 2 of the prototype, Fig. 3 presents an overview of the prototype processes, their inputs and outputs and their interoperability. Each one of these five prototype processes is described in more detail in the following sub-sections.

1. Using as inputs the URDR that contains existing ALROs, as well as papers, RAs and RS, the ALR radius computation engine computes the LCR index.
2. Next, using as inputs the ALR Corpus and the training models built by selected researchers, MLM provide the ALR learning model used by the Multilevel-based Relevant ALR Corpus. MLM also enrich the ALR Corpus to provide the ALRO.
3. The Multilevel-based Relevant ALR Corpus computes the DTb-index that measures the relevancy of each paper in the ALR corpus.
4. Making use of the generated and enriched ALRO, the ALR Refine & Recommendation engine suggests the Paper References list to the researcher.

4.1 ALR radius computation

ALR radius computation is used to rank the relevancy of papers to be included in the ALR, according to the researcher selection (RS) and researcher annotations (RAs). Computation of the LCR index is defined as a sub-algorithm of the semantic ALR selection search that identifies the ALR corpus according to the RS and RAs. Here, selection metadata and selection parameters may be used interchangeable. To identify an ALR corpus as shown in the Step 1 of Fig. 4, the selection parameters are classified into three categories:

1. The LCR index is computed based on the TDM approach.
2. The document's Researcher Annotations (RAs) consist of: key Findings, free Tags, personal Notes and pre-defined Tags.
3. In sort-based selection, a specified number of documents are sorted according to a particular order. For an ALR in a given field, the researcher may need to keep:

a. Z% of relevant documents that are X years old or less, and
b. (100-Z)% that are more than X years old.

In the following paragraphs, the TDM semantic topic search is explained in detail.

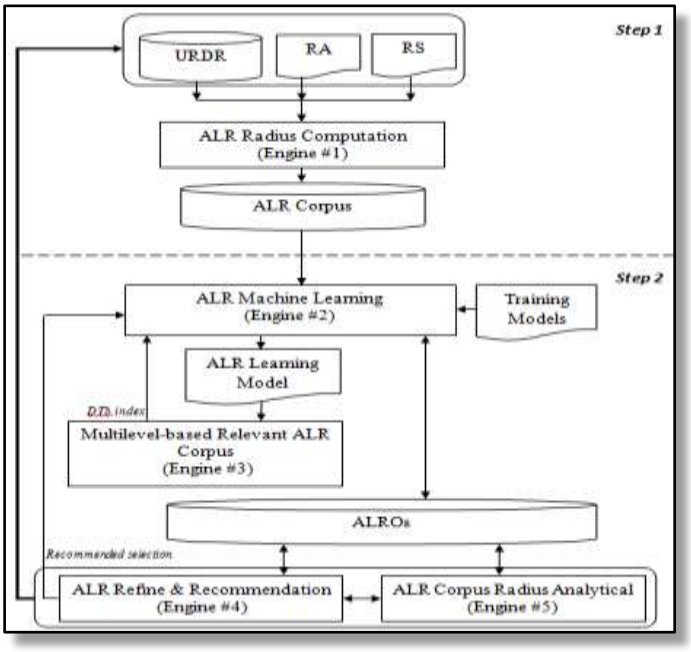

Figure 3: Interoperability of the prototype processes

A. Discipline and language researcher selections step

In step A, the volume of documents to be considered for the rest of the process may be reduced, based on:

1. Discipline selection: selecting all documents that are in the Meta Corpus of a given discipline, e.g., Biology and Computer Science.
2. Language selection: limiting the documents to be considered for the ALR to a specific language; the default value is English.

The selection query uses the document metadata in the URDR.

Let DC be the chosen discipline, let LG be the given language, let DISCIPLINE be the metadata that records the discipline of the documents in URDR, let LANGUAGE be the metadata that records the language of the documents in URDR and let DiscLan_Corpus (DC,LG) be the set of documents in the language LG that are in the discipline DC.

DiscLan_Corpus (DC, LG) is obtained as follows:

*DiscLan_Corpus (DC, LG) = [**select in** URDR the Documents **where***
*DISCIPLINE is "DC" **and***
LANGUAGE is "LG"]

This query to the URDR extracts only papers in the specified discipline and language.

Let C_1 be the corpus of papers obtained in step A.

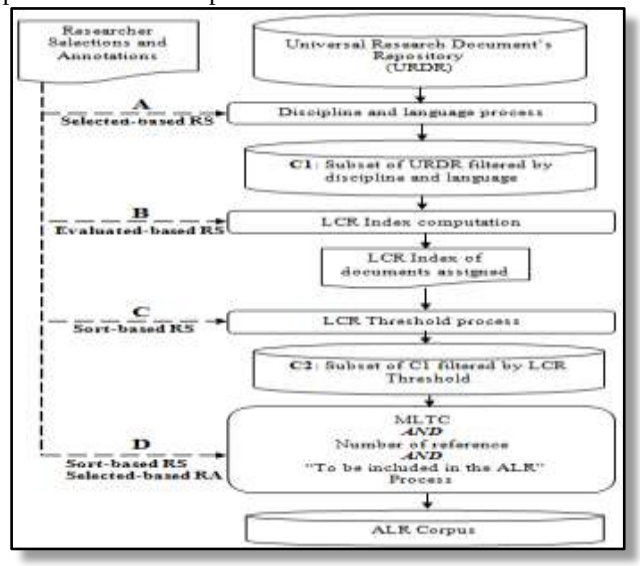

Figure 4: Steps in a semantic ALR selection search

B. LCR index computation step

Using the set of papers extracted in step A, the LCR index is computed next in step B based on the evaluation-based selections: main topic, keyword, title and description.

The impact of each of these selections is computed to identify the papers that best match the researcher selections:

1. First, the similarity matching of each evaluation-based selection with a predefined selection of papers is evaluated within the range [0,1]: 1 means the most similar while 0 means the least similar.
2. Next, based on their predefined weight and the similarity matching value, the LCR index is computed.

The LCR index computation step consists of five sub-steps, a to e.

a. Similarity matching of researcher main topic with topics extracted from document abstracts

The similarity matching of the researcher main topic with the topics extracted from the document abstracts is first computed using the topic detection ML model called BM-Scalable Annotation-based Topic Detection (BM-SATD) [38]. More specifically, BM-SATD uses multiple relations within a term graph and detects topics from the graph using a graph analytical method. BM-SATD combines semantic relations between terms with co-occurrence relations across the document, by making use of the document annotations.

Here, the similarity matching is based on the n-gram approach where the value n is used as the weight [39]: when the i-gram expression of the researcher main topic is found in the abstract, the weight i is associated with this expression.

b. Similarity matching of researcher keywords with document keywords

The similarity matching of the researcher keywords is computed next by making use of the KEYWORDS sections of the documents. The impact value is the number of researcher selection keywords that are similar to the KEYWORDS section.

c. Similarity matching of researcher title with document titles

Before this similarity matching computation, the researcher title and document titles are pre-processed to filter noise. This consists in stemming, phrase extraction, part-of-speech filtering and removal of stop-words. Next, based on the terms obtained, the maximum n-gram of the researcher title which is met in the document title is used as the title selection impact value.

d. Similarity matching of researcher research topic description with document abstracts

The researcher research topic description is semantically compared with the document abstract in order to measure the semantic similarity level. This similarity matching makes use of WordNet::Similarity, which applies six measures of similarity and three measures of relatedness; thus, several terms may be semantically the same. To measure this similarity, the TF-IDF approach is extended to meet our objective by applying it to the vocabulary of the corpus instead of the document itself.

e. LCR index computation

Finally, when the similarity matching of each evaluation-based selection has been completed through sub-steps a to d, the LCR index within the [0,1] range can be computed. Note that the LCR index is a weighted sum of the computed value of each evaluation-based selection.

The difference in weight between two consecutive evaluation-based selections (i.e., selection i and selection i+1) is a predefined constant value.

C. Literature Corpus Radius (LCR) threshold selection step

In this step, a set of documents is sorted or selected according LCR index value. For example, a researcher may indicate that the LCR threshold is 0.7; the output will then be a subset of corpus C whose LCR index is greater than or equal to 0.7. When the researcher does not give this selection, the set of documents obtained in step A above (Discipline and language researcher selections) is used as the input of this step.

Let C_2 be the corpus of documents obtained in step C.

D. MLTC AND Number of references AND "To be included in the ALR" step

MLTC is the Mix Literature Temporal Coverage. Let MLTC (x, y) with its number of selections equal N: this means the researcher expects to have at most N documents, with a maximum of (100-x)% (i.e., $\frac{N}{100} \times (100 - x)$) that are at most y years old, and including all the documents tagged "To be included in the ALR". Note that the latter documents have priority.

First, a list (in descending order) is created based on the LCR index applied to corpus C1 where the documents tagged "To be included in the ALR" are at the top due to their priority.

Let All_C1 be this list. New_C1 is defined as a sub-list of C1 in which the document age is less than or equal to y, and Old_C1 contains documents older than y.

Let $A = \frac{N}{100} \times x$ be the length of New_C1 and $B = \frac{N}{100} \times (100 - x)$ be the length of Old_C1. To take into account the three selections made in sub-step D.

Note that, when the number of documents in All_C1 is less than N, all the documents are considered affinity matches for the ALR; in that case, the MLTC selection is ignored.
However, when there are not enough documents whose age is less than or equal to y to satisfy the MLTC selection, a new MLTC is provided in order to reach the number A. But if the researcher requires the MLTC selection to be met, some documents are removed from New_C1 in order to meet the selected MLTC (x, y).
If an "OR" has been placed between the researcher selections, the LR corpus will be defined as the union of the C2 subsets provided by the MLTC process, the Number of references process and the "To be included in the ALR" tags.

4.2 ALR Machine Learning

ALR Machine Learning for semantic ALR selection is the core of the prototype. It is the only process that interacts with all the algorithms of the other MLM, combining the TDM and MLM approaches to discover hidden information in papers. ALR machine learning is a supervised MLM that makes use of a training set in order to provide the learning model composed of three sub-models: section recognition learning model, citation-based learning model and text-based learning model.

4.3 Multilevel-based relevant ALR Corpus

The multilevel-based relevant ALR Corpus is presented here. It is used to evaluate the relevancy of a paper based on a number of scientometric measurements. Here, relevancy is not based on RAs and RS; instead, the input corpus used by the multilevel-based relevant ALR Corpus is the ALR Corpus obtained through the ALR's semantic search based on RAs and RS. Three types of ALR Index are defined in the prototype: personal, collaborative and dynamic topic-based (DTb).

With the personal index, the ALR can be restricted to documents tagged by the researcher as "To be included in the ALR". The collaborative index extends the personal index by including documents tagged "To be included in the ALR" by a specific community of researchers.

The dynamic topic-based index (DTb index) selects documents for the ALR when the researcher has not requested a personal or collaborative index. The DTb index is a weighted sum of the values that denote the importance of the different inputs considered:

1. Key findings and peer citations index
2. Venue index
3. Document references index
4. Authors and their affiliated institutes

4.4 ALR Refine & Recommendation MLM

The ALR Refine & Recommendation MLM is presented here. The input is the ALR Corpus of relevant and enriched papers identified automatically by the prototype to recommend selections parameters to a researcher (see previous sections). This MLM may next recommend three different aspects of the ALR selection (Fig. 5):

1. The list of papers to be included in or removed from the ALR
2. The number of references (i.e., papers) to be considered for the ALR
3. The % of Mix Literature Temporal Coverage (MTLC) to be included in the list of references

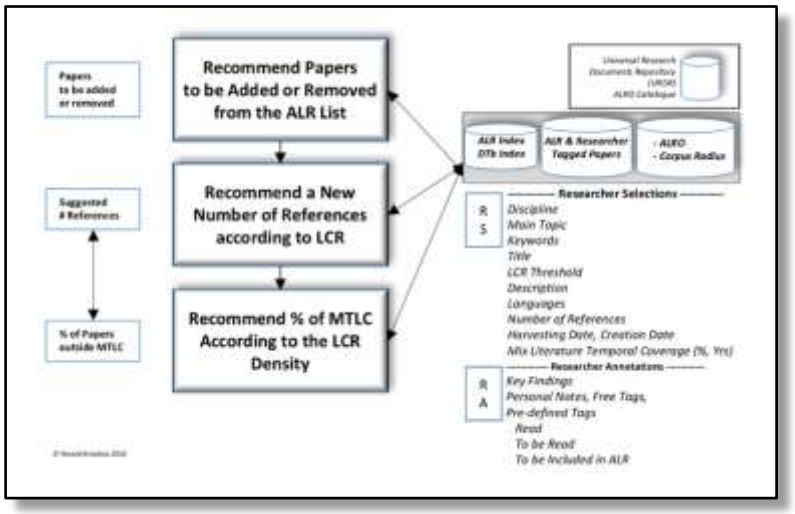

Figure 5: Refinement & Recommendation MLM

To help the researcher to choose the right combination of parameters (RS), the refinement function makes recommendations in the following three areas:

1. Identification of documents to form the recommended list for the ALR.
2. Identification of the optimal number of documents as references to include in the ALRO. This recommendation is related to the LCR and based on the most relevant documents closest to the selected topic; the highest number will be the proposed number of references. The sub-steps are:
3. Identification of the % of MTLC to be part of the ALR.

4.5 ALR Corpus Radius Analytics

The ALR Corpus Radius Analytics presents a number of ways of viewing the list of documents for drill-down purposes. This sub-section describes the concepts used in producing an assisted ALR, including:

1. The Timeline of a Document-based Literature Corpus Radius
2. The Literature Corpus Radius (LCR)

Two classes of documents are defined: citing documents and cited documents. For a better understanding, let d be a considered document; a citing document is a document that cites document d while a cited document is a document that is cited by document d.

V. Prototype Performance Evaluation Through Simulations

This section presents an evaluation of the performance of the prototype through a number of simulations limited to the identification of relevant papers for an ALR.

5.1 Datasets

Two datasets were used for the simulations:

1. A dataset harvested from existing scientific papers repositories
2. A baseline dataset

5.1.1 Dataset harvested from databases

For the simulations, 5,000 scientific papers were harvested from scientific paper search engines such as ReseachGate, Academia, ScienceDirect, Scopus, Google scholar, Citeseerx and IEEE Xplore. The papers dealt with various research topics of computer sciences such as Software, Hardware and Architecture, Human-Computer Interaction and Computer Science Applications. In the context of these simulations, the research topics are treated as domains. The other metadata were collected as bibliographic references.

For each paper, the downloaded bibliographic files were parsed to extract the metadata. A scenario was defined as a set of two simulator runs, one on each domain dataset. For the simulation, the metadata of one paper in the dataset (discipline, language, title, topic, keywords and abstract) were used as the RS and RA parameters.

5.1.2 Baseline dataset

For the present study, we had already produced a manual ALR that included all the papers listed in our References section. This manually assembled list was used as the baseline dataset to evaluate the performance of the prototype. The baseline dataset consisted of 58 papers dealing with both general and specific topics within the domain. Here, a scenario was defined as one simulator run where the 39 papers constituted the dataset. The metadata of the present study were used as the RS and RA parameters.

5.2 Performance criteria

The prototype was evaluated from the viewpoint of the researchers. As in [12], two performance criteria were used to assess the relevancy of the papers in terms of accuracy. The accuracy is defined as the percentage of true classifications.

Considering the sets of relevant papers (REL) and non-relevant papers, (NREL), true relevant (TR) denotes the papers classified as REL when they really are, while false relevant (FR) denote the papers classified as REL when they are not. Thus, with the same logic, the papers classified as NREL can be true non-relevant (TN) or false non-relevant (FN). For each type of dataset, the definition of a scenario is given in sections 5.1.1 and 5.1.2 according to the type of dataset. Accuracy, denoted by a, was computed as follows for each scenario:

$$a = \frac{TR + TN}{TR + FR + TN + FN}$$

To identify TR, FR, TN and FN for each scenario, a target paper was chosen for the domain; next, the metadata of this target paper were used as the researcher selection parameters and the references papers in the output set of the prototypes were compared to the cited papers of the target paper. Through this comparison, TR, FR, TN and FN were defined. Let ai,j be the accuracy of the scenario ith of the dataset j; the average accuracy is defined as follows:

$$Avg\_a_i = \frac{\sum_{j=1}^{D} a_{i,j}}{D}$$

where D denotes the number of datasets.

5.3 Related ranking approaches for comparison purposes

There are two other works on scientific paper ranking:

- PTRA [6]
- ID3 [12].

PTRA and ID3 are described in section 2.1. Table 2 presents a summary of the criteria taken into account by each ranking approach: the bottom line of Table 2 lists all the criteria used in the prototype ranking approach. The performance of the prototype approach was compared against the performance of PTRA [6] and ID3 [12] on the same datasets and scenarios. It is observed that for ranking a cited document as relevant, the prototype considers more criteria, such as venue age, citation category, authors' impact, etc.

Table 2: Criteria taken into account in three paper ranking approaches

Approaches	Year of publication	Citation number	Reference	Venue type	Venue age	Authors' impact	Citation category	Venue impact	Authors' institutes	Citing document of cited document
PTRA [6]	X	X		X						
ID3 [12]	X	X	X	X						
ALR prototype	X	X	X	X	X	X	X	X	X	X

5.4 Analysis of the simulation results

This section presents the analysis of the simulation results in terms of papers' relevancy for the two datasets.

5.4.1 Simulation using the dataset harvested from databases

Fig. 6 shows the average accuracy for the three different simulations (prototype, ID3 and PTRA). The horizontal axis represents the sequence number of the simulation scenarios and the vertical axis represents the average accuracy of the associated scenario.

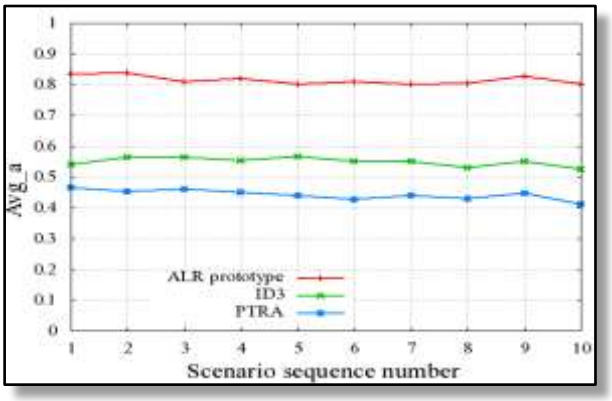

Figure 6: Average accuracy vs Scenario sequence number

It is observed that the prototype (in red) performs better than ID3 (in green) and PTRA (in blue): prototype has an average accuracy of 0.82 per scenario while ID3 has an average of 0.54 per scenario. The average relative improvement in accuracy (defined as [Avg_a of the prototype R – Avg_a of ID3]) of the prototype in comparison to ID3 is 0.28 (28%) per scenario. The prototype outperformed ID3 and PTRA. This performance might be attributable to the use of additional bibliometric metadata.

5.4.2 Simulation using the baseline dataset

Table 3 presents the accuracy when the list of papers in the baseline dataset (i.e., the references cited in this paper) is used as the dataset for simulations. It produced an average accuracy (Avg_a) of 78.17% while ID3

produced an accuracy of 53.98%. The relative improvement in accuracy of the prototype as compared to ID3 is 24.19%.

Note that all the simulations are based on limited datasets, and should be extended later to larger datasets.

Table 3: Summary of performance criteria (accuracy) using the baseline dataset

Approaches	Avg_a (%)
PTRA (Hasson et al., 2014)	41.34
ID3 (Rúbio & Gulo, 2016)	53.98
ALR prototype	78.17

VI. Summary And Future Work

With the evolving, interdisciplinary nature of research and online access to research papers, there is a need to facilitate the iterative process of building a corpus for an assisted literature review (ALR). The aim of the present study is to assist researchers in finding, evaluating and annotating relevant papers, and to make them available at any time in an iterative process.

This paper has proposed an ALR prototype based on machine learning model to identify, rank and recommend relevant papers for an ALR. Using TDM models, MLM and a classification model that learns from researchers' annotated data (RA) and semantic enriched metadata, The prototype assists in identifying and recommending papers that meet a researcher selection (RS) of parameters, including specific topic, title, language, discipline, papers age, number of references and other metadata.

This paper has presented TDM models, related MLM and an enhanced metadata ecosystem that can help researchers produce ALRs. These include:

1. MLM designed to semantically harvest a Universal Research Documents Repository (URDR) according to a researcher selection;
2. Literature Corpus Radius (LCR) MLM, which compute the distance from each paper to the center of the Literature Corpus defined by the researcher selection for a specific topic, concept or area of research;

The performance of the prototype has been evaluated through a comparison against a baseline manual LR using a number of simulations. In terms of accuracy, the prototype provided an average accuracy of 0.82 per scenario while ID3 provided an average of 0.54 per scenario. In comparison to ID3, the prototype yielded an average relative improvement in accuracy of 28% per scenario.

The areas of future work on the prototype will be:

1. Abstract of Abstracts summarization (AoA): AoA for scientific papers will be an extension of this prototype; more specifically, abstracts will be used as input for our scientific paper summarization technique to generate the AoA.
2. Digital Resources Metadata Enrichment (DRME): the next prototype will implement a new semantic discovery tool called DRME to help aggregate metadata from papers that have not published their metadata.

References

[1] Mayr P, Scharnhorst A, Larsen B, Schaer P, Mutschke P (2014) Bibliometric-Enhanced Information Retrieval. Paper presented at the 36th European Conference on IR Research (ECIR), Amsterdam, The Netherlands, 13-16 April 2014

[2] Carlos ASJG, Thiago RPMR (2015) Text Mining Scientific Articles using the R Language. Paper presented at the 10th Doctoral Symposium in Informatics Engineering, Porto, Portugal, 29-30 Jan. 2015

[3] Gulo CASJ, Rubio TRPM, Tabassum S, Prado SGD (2015) Mining Scientific Articles Powered by Machine Learning Techniques. OASIcs-OpenAccess Series in Informatics 49:21-28. doi:http://dx.doi.org/10.4230/OASIcs.ICCSW.2015.21

[4] Zhang M, Zhang X, Hu Y (2015) Ranking of Collaborative Research Teams Based on Social Network Analysis and Bibliometrics. Paper presented at the 12th International Conference on Cooperative Design, Visualization, and Engineering (CDVE), Mallorca, Spain, 20-23 Sep. 2015

[5] Madani F, Weber C (2016) The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis. World Patent Information 46:32-48. doi:http://dx.doi.org/10.1016/j.wpi.2016.05.008

[6] Hasson MA, Lu SF, Hassoon BA (2014) Scientific Research Paper Ranking Algorithm PTRA: A Tradeoff between Time and Citation Network. Applied Mechanics and Materials 551:603-611. doi:http://dx.doi.org/10.4028/www.scientific.net/AMM.551.603

[7] Beel J, Langer S, Genzmehr M, Gipp B, Breitinger C, Nurnberger A (2013) Research paper recommender system evaluation: a quantitative literature survey. Paper presented at the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, Hong Kong, China, 12th October 2013

[8] Cataldi M, Di Caro L, Schifanella C (2016) Ranking Researchers Through Collaboration Pattern Analysis. Paper presented at the European Conference on Machine Learning and Knowledge Discovery in Databases, Riva del Garda, Italy, 19-23 September 2016

[9] Franceschini F, Maisano D, Mastrogiacomo L (2015) Influence of omitted citations on the bibliometric statistics of the major Manufacturing journals. Scientometrics 103 (3):1083-1122. doi:http://dx.doi.org/10.1007/s11192-015-1583-9

[10] Wang S, Xie S, Zhang X, Li Z, Yu PS, Shu X (2014) Future Influence Ranking of Scientific Literature. Paper presented at the Society for Industrial and Applied Mathematics (SIAM) International Conference on Data Mining, Philadelphia, Pennsylvania, USA, 24-26 Apr. 2014

[11] MASIC I, BEGIC E (2016) Evaluation of Scientific Journal Validity, It's Articles and Their Authors. Stud Health Technol Inform 226:9-14. doi:http://dx.doi.org/10.3233/978-161499-664-4-93-5

[12] Rúbio TRPM, Gulo CASJ (2016) Enhancing Academic Literature Review through Relevance Recommendation. Paper presented at the 11th Iberian Conference on Information Systems and Technologies, Gran Canaria, Canary Islands, Spain, 15 - 18 Jun 2016

[13] Marx W, Bornmann L (2016) Change of perspective: bibliometrics from the point of view of cited references—a literature overview on approaches to the evaluation of cited references in bibliometrics. Scientometrics 109 (2):1397-1415. doi:http://dx.doi.org/10.1007/s11192-016-2111-2

[14] Dong Y, Johnson RA, Chawla NV (2016) Can Scientific Impact Be Predicted? IEEE Transactions on Big Data 2 (1):18-30. doi:http://dx.doi.org/10.1109/TBDATA.2016.2521657

[15] Bornmann L, Stefaner M, Anegón FdM, Mutz R (2014) Ranking and mapping of universities and research-focused institutions worldwide based on highly-cited papers: A visualisation of results from multi-level models. Online Information Review 38 (1):43-58. doi:http://dx.doi.org/doi:10.1108/OIR-12-2012-0214

[16] Packalen M, Bhattacharya J (2015) Neophilia Ranking of Scientific Journals. National Bureau of Economic Research Working Paper Series 21579. doi:http://dx.doi.org/10.3386/w21579

[17] Bornmann L, Stefaner M, Anegón FdM, Mutz R (2015) Ranking and mappping of universities and research-focused institutions worldwide: The third release of excellencemapping.net. COLLNET Journal of Scientometrics and Information Management 9 (1):65-72. doi:http://dx.doi.org/10.1080/09737766.2015.1027090

[18] Wan X, Liu F (2014) WL-index: Leveraging citation mention number to quantify an individual's scientific impact. Journal of the Association for Information Science and Technology 65 (12):2330-1643. doi:http://dx.doi.org/10.1002/asi.23151

[19] Zhang R, Li W, Liu N, Gao D (2016) Coherent narrative summarization with a cognitive model. Computer Speech & Language 35:134-160. doi:http://dx.doi.org/10.1016/j.csl.2015.07.004

[20] Carenini G, Cheung JCK, Pauls A (2013) MULTI-DOCUMENT SUMMARIZATION OF EVALUATIVE TEXT. Computational Intelligence 29 (4):545-576. doi:http://dx.doi.org/10.1111/j.1467-8640.2012.00417.x

[21] Dunne C, Shneiderman B, Gove R, Klavans J, Dorr B (2012) Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. Journal of the American Society for Information Science and Technology 63 (12):2351-2369. doi:http://dx.doi.org/10.1002/asi.22652

[22] Mohammad S, Dorr B, Egan M, Hassan A, Muthukrishan P, Qazvinian V, Radev D, Zajic D (2009) Using citations to generate surveys of scientific paradigms. Paper presented at the Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Boulder, Colorado, USA, 31 May - 5 June 2009

[23] Conroy JM, Davis ST (2015) Vector Space and Language Models for Scientific Document Summarization. Paper presented at the Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies, Denver, Colorado, USA, 31 May – 5 Jun. 2015

[24] Dyas-Correia S, Alexopoulos M (2014) Text and Data Mining: Searching for Buried Treasures. Serials Review 40 (3):210-216. doi:http://dx.doi.org/10.1080/00987913.2014.950041

[25] Ronzano F, Saggion H (2016) An Empirical Assessment of Citation Information in Scientific Summarization. Paper presented at the 21st International Conference on Applications of Natural Language to Information Systems (NLDB), Salford, UK, 22-24 Jun. 2016

[26] Widyantoro DH, Amin I (2014) Citation sentence identification and classification for related work summarization. Paper presented at the International Conference on Advanced Computer Science and Information Systems (ICACSIS), Jakarta, Indonesia, 18-19 Oct. 2014

[27] Pedram VA, Omid SS (2015) Scientific Documents Clustering Based on Text Summarization. International Journal of Electrical and Computer Engineering (IJECE) 5 (4):782-787

[28] Huang S, Wan X (2013) AKMiner: Domain-Specific Knowledge Graph Mining from Academic Literatures. Paper presented at the 14th International Conference on Web Information Systems Engineering (WISE), Nanjing, China, 13-15 Oct. 2013

[29] Caragea C, Bulgarov F, Godea A, Das Gollapalli S (2014) Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach. Paper presented at the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25-29 Oct. 2014

[30] Chen J, Zhuge H (2014) Summarization of scientific documents by detecting common facts in citations. Future Generation Computer Systems 32:246-252. doi:http://dx.doi.org/10.1016/j.future.2013.07.018

[31] Jaidka K, Khoo CSG, Na JC (2010) Imitating Human Literature Review Writing: An Approach to Multi-document Summarization. Paper presented at the 12th International Conference on Asia-Pacific Digital Libraries (ICADL), Gold Coast, Australia, 21-25 Jun. 2010

[32] Yeloglu O, Milios E, Zincir-Heywood N (2011) Multi-document summarization of scientific corpora. Paper presented at the ACM Symposium on Applied Computing (SAC), TaiChung, Taiwan, 21-24 Mar. 2011

[33] Jaidka K, Khoo CSG, Na JC (2013) Literature review writing: how information is selected and transformed. Aslib Proceedings 65 (3):303-325. doi:http://dx.doi.org/doi:10.1108/00012531311330665

[34] Agarwal N, Gvr K, Reddy RS, Rose CP (2011) SciSumm: a multi-document summarization system for scientific articles. Paper presented at the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations, Portland, Oregon, USA, 19-24 Jun. 2011

[35] Patil SR, Mahajan SM (2012) Scientific Research Paper Summarization On The Basis Of Research Relevant Term Identification. Paper presented at the International Conference and workshop on Emerging Trends in Technology (ICWET), Mumbai, Maharashtra, India, 24-25 Feb. 2012

[36] Jaidka K, Khoo CSG, Na JC (2013) Deconstructing human literature reviews–a framework for multi-document summarization. Paper presented at the 14th European Workshop on Natural Language Generation, Sofia, Bulgaria, 8-9 Aug. 2013

[37] Bechhofer S, Buchan I, De Roure D, Missier P, Ainsworth J, Bhagat J, Couch P, Cruickshank D, Delderfield M, Dunlop I, Gamble M, Michaelides D, Owen S, Newman D, Sufi S, Goble C (2013) Why linked data is not enough for scientists. Future Generation Computer Systems 29 (2):599-611. doi:http://dx.doi.org/10.1016/j.future.2011.08.004

[38] Brisebois R, Abran A, Nadembega A, N'techobo P (2017) A Semantic Metadata Enrichment Software Ecosystem based on Sentiment and Emotion Metadata Enrichments. International Journal of Scientific Research in Science Engineering and Technology (IJSRSET) 03 (02):625-641

[39] Bertin M, Atanassova I, Sugimoto CR, Lariviere V (2016) The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. Scientometrics 109 (3):1417-1434. doi:10.1007/s11192-016-2134-8

* Ronald Brisebois " An Assisted Literature Review using Machine Learning Models to Identify and Build a Literature Corpus " International Journal of Engineering Science Invention (IJESI) 6.7 (2017): 72-84.