A CLASSIFICATION MODEL FOR MINING RESEARCH PUBLICATIONS FROM CROWDSOURCED DATA

Omisore M. O.
Federal University of Technology Akure
PMB 704, Akure, Ondo State, Nigeria
+234-703-1967847
ootsorewilly@gmail.com

ABSTRACT
Automatic access of natural language meaning is a prominent way of implementing search engines for document classification. The technique is difficult and often presents search results in rough approximates. It has minimal linguistic processing performed to identify content words like nouns and verbs in indexed documents. However, word frequency in documents can be taken as clues to their similarity. In this paper, a research problem on how to formulate classification model for document is proposed. Techniques for analyzing text contents for the purpose of document clustering and content modeling are focused as objectives. The abstracts look more at modeling textual entailment recognition to determine the semantic relationship in pool of crowdsourced documents on probabilistic setting. The prospective innovation is to improvise a rich and novel mining approach base on semantic relationship of research documents.

Categories and Subject Descriptors
H.3.2 **[Information Storage]**: Record classification; H.3.4 **[Information Storage and Retrieval]**: Content Analysis and Indexing; H.3.7 **[Digital Libraries]:** Document Classification

General Terms
Design; Algorithms

Keywords: Document Mining and Classification, Natural Language Processing, Entailment Recognition, Crowdsourcing.

1. RESEARCH PROBLEM
The persistent interest in study of science or social study of science is a consequence of the reflexive endeavor to comprehend and assimilate science and the growth of scientific knowledge perhaps, together with policy intentions to design evaluation and stimulus mechanisms [1]. This is further enhanced by importance associated with science as the driver of social and economic progress in all domains.

Document classification is the task of sorting a set of documents into categories or classes based on some predefined subject matters or topics by construing texts inherent to the documents. This is a cross-pollinated task between information retrieval (IR) and machine learning fields. Text categorization is required in IR because a user queries for information in a Web search engine to be filed according to relevance to the query and its contents. Text categorization also holds a great interest for researchers in the field of machine learning as it provides an excellent benchmark for their own techniques and methodologies [2].

Document classification is a problem in information science that involves manual or algorithmic assignment of document to one or more classes usually because of their contents. Although, manual method is a province of library science where human experts are engaged to perform the classification tasks however, algorithmic approach is mainly used in information science by utilizing unsupervised or supervised techniques to classify resources [3]. Previously, majority of works in this domain focused on factual information and text categorization thereby classifying documents according to subject matters such as politics or sports. Question answering systems aimed to answer fact-oriented questions such as who did what, when and where?. Similarly, research in text summarization mainly focused on collections of news articles where it is important to keep analysis of figures in facts [4].

Crowdsourcing is a process of obtaining needed services, ideas, or content by soliciting contributions from a large group of people (online community) rather than via physical contacts [5]. This process is often used to subdivide tedious work and has been successfully applied in different areas of Natural Language Processing [6]. It combines efforts of many volunteers who contribute their initiative in respective proportions to form a greater result. Crowdsourcers are primarily motivated by benefits such as ability to get large amount of information at a relatively inexpensive cost [7]. However, a major consequence of the vast growth of the Internet is the enormous amount of electronic data, mostly in form of text [8] thereby making it very difficult for mining relative interlinked documents for research purpose.

Justifications from above shows there is need to develop a model for mining and creating interlinks in research publications that are crawled or crowdsourced from one or multiple scientific data sources. Hence, this paper endeavors to improvise by proposing a model for mining and interlinking research publications from crowdsourced data. The remaining part of this paper is organized as follows: Section 2 presents justification for the significance of the research problem; Section 3 summarizes the state of the art of the problem and current knowledge of the problem domain; Section 4 presents the preliminary research plan; Section 5 sketches the research methodology conceptualized to be applied; Lastly, Section 6 describes the expected contributions of the study and the novelty of the proposal.

2. SIGNIFICANT PROBLEMS
In the last two decades, there has being an increasing need for very efficient content-based document management system which does not just include document classification, alternatively known as text categorization, but as well covers the process of labeling and annotating documents for mining and interlinking research articles from different categories of a predefined set [9]. The Internet which had emerged as an important information resource has become a central challenge in this context [8]. In order to satisfy human needs in this regard, development of an intelligent system for processing textual linguistics, a view supported by vision of Semantic Web, deserves special attention. Yet, current technologies such as search engines or prototypical question answering systems typically consider semantic information only in limited ways. This can be easily observed by the broad number

of digital libraries on the Web [10]. Like the case of physical libraries, digital libraries also suffer some difficulties in organizing and searching for information [11]. IR systems have contributed significantly to management of textual information but their success still depend on fast and vast algorithm that can efficiently process textual content in documents [12]. This is simply because human expert does not have to be solely responsible for capturing new information from data sources, indexing and cataloguing the information else IR systems will remain tedious and time consuming.

In many real life scenarios, classifying a large amount of archival documents such as research publications, newspaper articles, and legal records has been a major concern until very recently when there had being proposals to develop systems that can classify documents into a fixed set of categories is highly desirable [13]. The efficiency, scalability and quality of document classification algorithms heavily rely on the representation of such documents. Among the set-theoretical, algebraic and probabilistic approaches, Vector Space Model (VSM), proposed in [14], describes how textual documents can be represented in vector spaces.

Natural Language Processing (NLP) offers powerful techniques for automatically classifying documents [48]. These techniques are predicated on the hypothesis that documents in different categories distinguish themselves by features of the natural language contained in each document. Salient features for document classification may include word structure, word frequency, and natural language structure in each document. A large portion of research in NLP seeks to better understand and process various types of information in text [15].

Attractive ways of solving complex tasks in NLP is to develop models that automatically learn by inductive or deductive training. Learning algorithms have traits to discover important information which can be manually encoded in rule-based systems [16]. In recent works, an acceptable level of accuracy has been achieved for natural language tasks like part-of-speech tagging, named entity recognition, word disambiguation and semantic role labeling [17].

Vector space model is an algebraic model of representing textual documents as vectors of identifiers such as index terms. The technique is used in information filtering and retrieval [18], indexing [19], and relevancy rankings [20]. In addition to efficiency attained in representing compact documents using VSM with reduced dimensions, the technique is as well capable of removing noise such as synonymy, polysemy or rare term use [17].

Despite the pros of VSM, long documents are poorly represented due to poor similarity values and also, search keywords must precisely match document terms else word substrings might result in a false positive match. Another great concern is the issue of semantic sensitivity where documents with similar context but varying vocabulary terms are not associated and thereby resulting in a false negative match.

Textual entailment is used to indicate the state in which the semantics of a natural language written text can be inferred from the semantics of another one. In theoretical and computational semantics, truth-conditional logic formalisms have been the standard framework for modeling natural language meaning over the last decades. Reasoning for NLP has been conceived as being more or less equivalent to logical reasoning [21]. For example, inferential relations between natural language sentences have been modeled in terms of their logical entailment.

Textual Entailment Recognition (TER) has been proposed recently as a generic task that captures major semantic inference needs across many sentiment-based applications [22]. Sentiment analysis which has attracted a great deal of attention as a major field of study is an NLP approach that deals with computational treatment of opinion and subjectivity in text. This subject area can be applied in question-answering and message filtering systems where sentiment information of texts are used to recognize and discard flames.

Textual entailment in semantic analysis is a directional relation between text fragments holding relation whenever the truth of one text fragment follows from another. In addition to the interesting characteristics of sentiment analysis, modeling inter-relations in their text collections which tend to be clean, edited, and grammatical in the past are now full of blog postings and user reviews. These incline to be less structured and sometimes contain incorrect spellings or grammatically incorrect. Textual Entailment Recognition (TER) can thereby focus on classification tasks where most of the problems are investigated in context of classifying a given document according to sentiment polarity (a two-class problem) or degree of positivity (a multi-class problem) expressed in the text. A form of sentiment summarization is then used in this context for detection of subjectivity by extracting the subjective portions of input documents.

In various TER challenges, several methods such as Lexical Relation, n-gram or subsequence overlap, syntactic matching, Semantic Role labeling, Logical Inference, Corpus statistics, and machine learning Classification have been applied. Most of these methods use some sort of lexical matching, such as logical inference for solving the text and hypothesis entailment problem.

Automatic access of natural language meaning is a difficult task. The most prominent search engines implementing document retrieval often roughly approximate it. Within such systems, minimal linguistic processing is performed mostly to identify content words like nouns and verbs in indexed documents. The frequencies of term occurrence are taken as clues to documents' meaning. In [23], a framework used to establish entailment relationship in texts: text (t), hypothesis (h) is proposed. Textual entailment is not the same as pure logical entailment with relaxed definition: "t entails h" ($t \Rightarrow h$) if, typically, a human reading t would infer that h is most likely true. The relation is directional because even if "t entails h", the reverse "h entails t" is much less certain. This follows [24] where entailment is referred to as relation between a pair of sentences with notion that the truth of a second sentence necessarily must follow that of the first one.

Despite the difficulties engaged in the application of supervised machine learning models to TER, such models have proved to be particularly successful in enhancing text classification basically, because textual entailment is an extremely complex natural language phenomenon [25, 26]. Generally, NLP tasks require a classifier to assign correct label to a target text fragment, looking at its context. In semantic role labeling observed in [27, 28], the goal was to assign correct roles to relevant text fragments with respect to a set of possible roles. Such fragments can be labeled using bag-of-word or syntactic interpretation models. However in contrast, TER requires processing two different texts between which complex syntactic-semantic relations holds and the goal is to classify such relations as either true or false entailment.

Moreover, typical bag-of-word models are not useful to capture the knowledge needed by the learning algorithms.

3. STATE OF ART OF THE PROBLEM

Entailment was a term derived from formal logic but now used as part of the study of semantics. All other essential semantic relations like equivalence and contradiction are defined in terms of entailment. Features of entailment can be utilized in two forms. The first form is when an analytic view of a sentence suffices and there is no need for prior information to validate the sentence truthfulness. For instance, ***my mother is a woman*** can quite be easily answered as '***true'***. Such can result in contradiction, when literal interpretation of the sentence answers false. For instance, ***my mother is a boy***. But in the second form, validating the sentences requires synthetic view and the validity of two different text fragments cannot be verified by their direct literal interpretations even with their dictionary meanings. Such validation would need some non-linguistic information about the text fragments to observe their relationship.

Early works on sentiment based classification of entire documents have often involved either the use of models inspired by cognitive linguistics or manual and semi-manual construction of discriminable word lexicons [29-31]. A methodology for real time sentiment extraction in finance domain was proposed in [32]. The system takes texts from web-based stock message boards and attempts to automatically label them as "buy", "sell" or "neutral". Unfortunately, the method constructs its word lexicon by manual selection and tagging of words from several thousand messages which ended up being indiscriminately slow.

Glickman [33] stated the primary task of IR for document mining as retrieving a set of documents relevant for an information need as specified by a search query, typically encoded in one or more natural language terms. Most users of document management systems, when formulating their retrieval query, usually employ terms that are expected to appear in relevant documents however, a document may not contain all query terms and yet be relevant.

Within application settings, a wide variety of techniques have been proposed to address semantic variability, ranging at different levels of representation and complexity in textual documents. This includes thesaurus-based term expansion [34] used technique for enhancing the recall of NLP systems and coping with lexical variability. Also, distributional similarity between words in documents has been an active research area for more than a decade. Qiu and Hans-Peter [35] present a probabilistic query expansion model based on a similarity thesaurus which was constructed automatically. The study shows that query expansion results in a notable improvement in document retrieval but, distributional similarity does not capture equivalence and entailment of meaning but rather broader meaning similarity.

Recently, lexical overlap shows a better accuracy in text comparison and modeling such technique to show textual entailment is a common problem in NLP tasks. Many heuristic techniques were applied within various applications to model such relations between text segments [36]. In the context of Multi-Document Summarization, [21] propose modeling the directional entailment between text fragments to identify redundant information. This baseline measure captures word overlap, using the statistical information of words that appear in both texts and weighs them based on their inverse document frequency.

The huge amount of information available online has given rise to personalization and filtering systems which constitute a specific type of information filtering technique that present items according to user's interests. In [11], a web-based personalized recommender system capable of providing learners with books that suit their reading abilities was developed. The research delivered an information filtering model that was experimented in a three-tier architecture with information of approximately 10000 books from different categories of computing kept in the end-tier (database) of the architecture. Fuzzy matching technique was implored to classify and rank books found suitable to learners base on such learners' reading abilities. The study proposed how the yokefellow cold-start problem inherent to Content Based systems can be assuaged by a cold starter component. The outcome tracked over a period of eight months shows that the model induces greater user satisfaction yet however, it did not test the quality of materials recommended and there was no provision for creating links between documents.

In various TER Challenges, several methods are applied on textual entailment task although most methods use some sort of lexical matching. Pazienza [37] defines a measure for textual entailment recognition based on the graph matching theory applied to syntactic graphs and this is used to estimate measure's parameters with Support Vector Machine. Also, Johan and Katja [38] combined shallow and deep NLP methods to tackle some textual entailment challenges. The shallow method was based on the established word overlap technique while a deep method using theorem proving, a machine learning technique, to combine features derived from both methods. The experiment was observed in two runs, one using all features from the two methods, yielding an accuracy of 0.5625, and the other using only the shallow feature and gives an accuracy of 0.5550.

Lexical-syntactic rules have been largely used in TER systems as they conveniently encode world knowledge into linguistic structures [25]. Given their importance, many methods, such as: [39, 40] have been proposed for the extraction of lexical-syntactic rules from large corpora. Unfortunately, these unsupervised methods in general produce rules that can hardly be used: noise and coverage are the most critical issues. Also, supervised approaches were demonstrated in [16], where lexical-syntactic rules were derived from examples in terms of complex relational features. This approach can easily miss some useful information and rules and can be unfortunately extremely misleading since it also derives similar decisions for different text.

Thus, this study will specifically look at the task of mining scientific research publications from crowdsourced sources that feature archives of unclassified documents and hence, require classification into specific categories base on the interlinks of their subject matters and other features. The study will observe investigating and implementing techniques that can be used to perform automatic classification or research articles.

4. RESEARCH PLANS

This proposal is aimed at developing a document classification model for mining research publications from crowdsourced documents in an online and real-time environment. The objectives include using NLP techniques for analyzing document similarity, document clustering and content modeling techniques to identify the similarity or closeness in mined documents.

To achieve the aim and objectives of the research, the following steps will be taken sequentially.

a. An extensive review of relevant literatures in IR and Mining, Document Management system, Natural Language Processing, Machine Learning, and Textual Entailing Recognition will be carried out with emphasis on deducting why, which and how classification of document is done.

b. A comprehensive study of existing research studies in the stated domains will be observed so as to possess a comprehensive knowledge about the state-of-the-art in aiding document classification with advances in Information Technology.

c. Existing repositories for scientific publication shall be explored for the purpose of crowdsourcing research articles from various fields. Based on this, a database model that can enhance management of such data will be designed for the purpose of experimental studies and evaluation.

d. A review of system results from the third TER challenge shall be presented and existing approaches to the task of modeling textual entailment will be selected. Also, models used in some baseline approaches shall be observed. These will include VSM, pair feature spaces, sentiment polarity and degree of positivity, and subjectivity and opinion detection.

e. A two-way versatile model that can recognize textual entailment between text fragments *text (t)* and *hypothesis (h)* will be designed. The recognition system will utilize lexical and syntactic features of the text fragments base on general probabilistic setting that formalizes a probabilistic notion of textual entailment proposed in (Dagan & Glickman, 2004). This is explained as:

i. Let T denote a space of possible texts, and t ∈ T is a specific text;
ii. Also, if H denotes the set of all possible hypotheses and hypothesis h∈H is a propositional statement which can be assigned a truth value.
iii. Then, a semantic state of affairs is captured by a mapping from H to {0=false, 1=true}, denoted by w: H→ {0, 1} where w is a word that represents concrete set of truth value assignments for all possible propositions. Accordingly, W denotes the set of all possible worlds.
iv. If probability of generating a true hypothesis h_t, which is related to a corresponding text, is determined by prior probability P(h), then the probability of t entailed in h is:

$$P(\mathrm{t} \Rightarrow \mathrm{h}) = \begin{cases} P(T_{r_h} = 1|t) > P(T_{r_h} = 1) & 1 \\ P(T_{r_h} = 1) + P(T_{r_h} = 0) = 1 & \text{otherwise} \end{cases} \quad (1)$$

where: T_{r_h} is the random variable whose value is truth value assigned to h in a given world.

v. Once knowing that t ⇒ h, then $P(T_{r_h} = 1|t)$ serves as probabilistic confidence value for h being true given t. Thus classification is based on lexical entailment probability of text information in the documents and it is given as:

$$P(T_{r_h} = 1|t) = \sum_{u \in h} max_{v \in t} P(T_{r_u} = 1|T_v) \quad (2)$$

where: $max_{v \in t} P(T_{r_u} = 1|T_v)$ is the maximum likelihood that $u, v \Rightarrow h$.

vi. Lastly, the documents are classified base on thier entailment scores (entscore) given as Eq. (3) and ranked with a predefined structured technique. Ranking documents was done with heuristic scores in [21].

$$entsore(t, h) = \frac{\sum_{w \in (t,h)} idf(w)}{\sum_{w \in h} idf(w)} \quad (3)$$

f. An experimental study will be carried out to validate the functionality and effectiveness of the proposed model. This will be implemented as a three tier architecture presented in Fig. 1. The model is conceptualized to operate in a web-enabled environment using Hyper-Text Markup Language (HTML) and Java Scripts for display and actions at front-end, Personal Home Page (PHP) scripts to coordinate middle-end activities, and My Structured Query Language Database Management System (DBMS) for back-end operations.

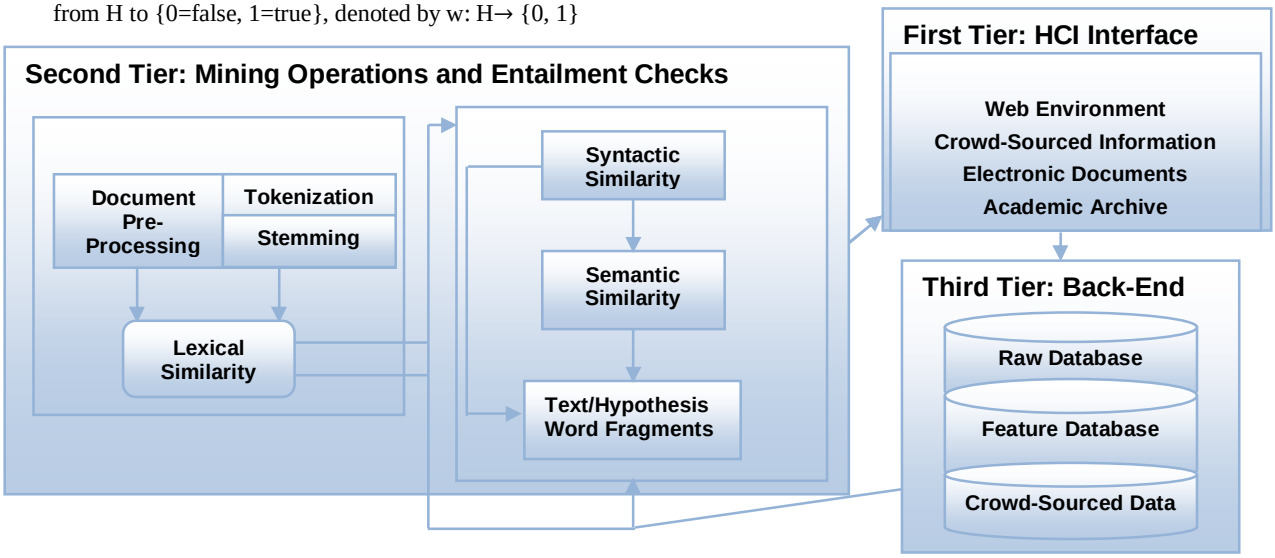

Fig 1: Architecture for Document Classification and Mining

g. Case study of crowdsourced research articles will be carried out on a computer system with Pentium IV processor at a speed of 2 GHz, 1GB RAM and 160 GB Hard Disk Space with Internet connection or WAMP suite or equivalent in the absence of proper Internet connection. The alternative can only fit in cases of computer systems running standalone web applications.

h. Result of the proposed model will be compared with results of other models such as machine learning-based model like Support Vector Machines, Maximum Entropy, and Naïve Bays, to evaluate the proposed model's accuracy and precision.

5. RESEARCH METHODOLOGY

A fundamental technology in current data mining and sentiment analysis applications is classification. This is because many problems of interest can be formulated as applying classification, regression, or ranking to given textual units. This research is envisioned to develop a model for mining research publications from different scientific data sources and creating interlinks that are used to automatically classify the documents base on their textual entailment.

The preliminary plans conceptualized to achieve vision in this proposal are to:

a. study key concepts in formulating common classification problems in sentiment analysis and document mining;

b. examine different linguistic models for learning textual entailment. Such models include: pair feature spaces model, two-way and three-way versatile models for textual entailment, sentiment polarity and degree of positivity and h-index or citation index.

c. develop a data centric model for crowdsourced documents;

d. design a broad model for mining documents in crowdsourced data centre. The mining algorithm will learn the entailments found in text pairs for the purpose of recognition and classification of documents;

e. implement the model design in (d) above in a web enabled environment for online and real-time access;

f. carry out an experimental study on the data centric crowdsourced documents in (c) above; and

g. evaluate the functionality of the proposed model with major focus on the TER capability and classification accuracy. Models based on h-index and citation index will be taken as benchmark in assessing the proposed model's performance.

6. EXPECTED CONTRIBUTIONS

This abstract describes the explorations of several textual mining concepts from a purely data-driven perspective, though it is believed that deeper linguistic knowledge can further improve the performance. The proposal states to design a textual entailment recognition model for classification and categorization of textual information. The functionality of the model will be evaluated by implementing the model design on a web platform study using crowd-sourced data.

This proposal seems to be innovative by adding to the research domain, a novel information and text mining approach to semantic enrichment of research documents in crowd (cloud). Also, interesting methods that can automatically classify publications according to established subject-based taxonomies like: Library of Congress classification, UNESCO thesaurus, or DOAJ subject classification; will be delivered. Therefore, new methods and model for connecting and interlinking scientific publications in digital libraries will not be isolated islands and that will, thereby, improve on the disadvantages of connecting publications using explicitly defined citations which are very restrictive.

In another innovative look, crowdsourcing can be used to annotate publications with richer metadata or to approve/disapprove annotations created using text-mining or other approaches. This study will look into models for semantically representing and annotating publications and application of crowdsourcing in the specialized domains of scientific publications, new methods, models and innovative approaches for measuring impact of publications since the most widely current used metrics for are based on citations and lastly, such new methods can be used for measuring performance of researchers.

7. ACKNOWLEDGMENTS

Special thanks to my supervisor and mentor, Dr. (Mrs.) Ojokoh B. A., of Computer Science department, Federal University of Technology, Akure, Ondo State, Nigeria.

8. REFERENCES

[1] Diana L. and Andrea S. 2012. *Mathematical Approaches to Modeling Science from an Algorithmic-Historiography Perspective*, Models of Science Dynamics, Understanding Complex Systems, Springer-Verlag Berlin, Heidelberg 2012, pp 23-66

[2] Morshed A. 2006. *Towards the Automatic Classification of Documents in User-generated Classifications*, Ph.D Thesis, Department of Information Technology and Communication, University of Trento, Trento, Italy

[3] Sebastiani, F. 2005. *Text Categorization*, In Alessandro Zanasi (ed). *Text Mining and its Applications*, WIT Press, Southampton, UK, 2005, pp 109-129.

[4] Pang B. 2006. *Automatic Analysis of Document Sentiment*, Ph.D Thesis, Computer Science Department, Faculty of Graduate School, Cornell University.

[5] Estellés-Arolas E. and González F. 2012. *Towards an Integrated Crowdsourcing Definition*, Journal of Information Science 38 (2): 189–200.

[6] Sarasua C., Simperl E., and Natalya F. 2012. *Crowdsourcing Ontology Alignment with Microtasks*, In Proceedings of the 11th International Conference on The Semantic Web Volume Part I, ISWC'12, pp. 525–541, Berlin, Heidelberg. Springer-Verlag

[7] Ebner W., Leimeister J., Krcmar H. 2009. *Community Engineering for Innovations: The Ideas Competition as a method to nurture a Virtual Community for Innovations*, R&D Management 39 (4): 342–356

[8] Bíró I. 2009. *Document Classification with Latent Dirichlet Allocation*, Ph.D. Thesis, Department of Information Sciences, Faculty of Informatics, Eötvös Loránd University.

[9] Omisore O. M. 2015. *Hybrid Recommendation Approach for Personalized Retrieval of Research Articles*, International Journal of Information Retrieval Research, 4(4), pp 42-60

[10] Lawrence S., Giles C. L., and Bollacker K. 1999. *Research Feature Digital Libraries and Autonomous Citation Indexing*, IEEE Computer, 32(6), pp. 67-71.

[11] Omisore M. O. and Samuel O. W. 2014. *Personalized Recommender System for Digital Libraries*, International Journal of Web-Based Learning and Teaching Technologies, volume 9, Issue 1, pp. 18-32.

[12] Ramos V. and Merelo J. 2002. *Self-Organized Stigmergic Document Maps: Environments as a Mechanism for Context Learning*, Proc. of the 1st Spanish Conference on Evolutionary and Bio-Inspired Algorithms, pp. 284-293.

[13] Hoe K., Lai W., and Tai T. 2002. *Homogeneous Ants for Web Document Similarity Modeling and Categorization*, Proceeding of the 3rd International Workshop on Ant Algorithms, Lecture Notes in Computer Science 2463, Springer-Verlag, pp. 256-261.

[14] Salton G.,Wong A., and Yang A. C. S. 1975. *A Vector Space Model for Automatic Indexing*, Communications of the Association of Computing Machinery, Vol. 18, pp. 229–237.

[15] Alekh A and Bhattacharyya P. 2005. Sentiment analysis: *A New Approach for Effective Use of Linguistic Knowledge and Exploiting Similarities in a Set of Documents to be Classified*, Proceedings of the International Conference on Natural Language Processing (ICON-05).

[16] Zanzotto F, Pennacchiotti M and Moschitti A. 2009. *A Machine Learning Approach to Textual Entailment Recognition*, Natural Language Engineering, 15(4): 551–82.

[17] Minnen, G., Carroll, J., and Pearce, D. 2010. *Applied Morphological Processing of English*, Natural Language Engineering, Vol. 7, Issue 3, pp. 207–223.

[18] Goodrum A. A. 2000. *Image Information Retrieval: An Overview of Current Research*, Informing Science, Vol 3(2), pp. 63-66

[19] Hanani U., Shapira B., and Shoval P. 2001. *Information Filtering: Overview of Issues, Research and Systems*, User Modeling and User-Adapted Interation, vol. 11, 203-259.

[20] Joran B., Bela G., Jan-Olaf S. 2009. *Information Retrieval on Mind Maps: What Could Be Good For*?, 5th International Conference on Collaborative Computing: Networking, Applications, and Workshop, Washington DC.

[21] Monz C and Maarten R. 2001. *Lightweight Entailment Checking for Computational Semantics*, In P. Blackburnan M. Kohlhase, editor, Proceedings ICoS-3.

[22] Dagan, I. and O. Glickman (2004. *Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability*, PASCAL Workshop on Learning Methods for Text Understanding and Mining.

[23] Quiñonero-Candela J., Dagan I., Magnini, B., and d'Alché-Buc, F. 2006. *Machine Learning Challenges*, Lecture Notes in Computer Science, vol. 3944, pp. 177-190, Springer.

[24] Crystal, D. 1998. *A Dictionary of Linguistics and Phonetics*, Oxford: Blackwell Publisher Ltd.

[25] Bar-Haim, R., Dagan, I., Dolan, B., and Magnini, I. 2006. *The Second Pascal Recognizing Textual Entailment Challenge*, In Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment, Venice, Italy.

[26] Giampiccolo, D., Magnini, B., Dagan, I., and Dolan, B. 2007. *The Third Pascal Recognizing Textual Entailment Challenge*, In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Prague.

[27] Gildea, D. and Jurafsky, D. 2002. *Automatic Labeling of Semantic Roles*, Computational Linguistics 28(3): 245–288.

[28] Carreras, X., and M`arquez, X. 2005. *Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling*, In Proceedings of the Ninth Conference on Computational Natural Language Learning, Ann Arbor, MI.

[29] Hearst M. 1992. *Direction-Based Text Interpretation as an Information Access Refinement*, In Paul Jacobs (eds), *Text-Based Intelligent Systems*. Lawrence Erlbaum Associates, pages 257-274.

[30] Huettner, A. and Subasic P. 2000. *Fuzzy Typing for Document Management*, In ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes, pp 26-27.

[31] Tong, R. M. 2001. *An Operational System for Detecting and Tracking Opinions in On-Line Discussion*, SIGIR Workshop on Operational Text Classification.

[32] Das S. and Chen M. 2001. *Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards*, In Proceedings of the Asia Pacific Finance Association Annual Conference.

[33] Glickman Oren. 2006. *Applied Textual Entailment*, Ph.D. Thesis, Department of Computer Science, Bar Ilan University, Ramat Gan, Israel

[34] Miller, G. A. 1995. *WordNet: A Lexical Databases for English*. Communications of the ACM, pp39-41.

[35] Qiu Y. and Hans-Peter F. (1993), *Concept Based Query Expansion*, Proceedings of the 16th Annual International Conference on Research and Development in Information Retrieval. Pittsburgh, PA, USA, pp 160-169.

[36] Geffet M and Dagan I. 2005. *The Distributional Inclusion Hypotheses and Lexical Entailment*, In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp 107-114, Ann Arbor, Michigan, USA

[37] Pazienza M. T., Pennacchiotti M., Zanzotto F. M. 2005. *Textual Entailment as Syntactic Graph Distance: a Rule Based and a SVM Based Approach*, Proceedings of 1st PASCAL Recognizing Textual Entailment Workshop.

[38] Johan B, Katja M. 2005. *Combining Shallow and Deep NLP Methods for Recognizing Textual Entailment*, Proceedings of 1st PASCAL Recognizing Textual Entailment Workshop.

[39] Lin D. and Pantel P. 2001. DIRT-Discovery of Inference Rules from Text, In Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD-01).

[40] Szpektor I. and Dagan I. 2008. Learning Entailment Rules for Unary Templates, In Proceedings of the 22nd International Conference on Computational Linguistics, Stroudsburg, PA, USA pp. 849-856.

[41] Bikel D. (2000), "A Statistical Model for Parsing and Word-Sense Disambiguation", Association for Computational Linguistics, Vol. 13, pp 155-163