A SOCIAL APPROACH FOR DOCUMENT ANALYSIS

Ibrahim Bounhas and Yahya Slimani
Department of Computer Science, Faculty of Sciences of Tunis, University of Tunis, 1060, Tunis, Tunisia

Keywords: Authority Study, Document Usage Type, Document Model, Logical Structure.

Abstract: This paper discusses the social constraints that should be taken into account in document analysis. In fact, a document is viewed and analysed as the object of a transaction between realizers and beneficiaries. Then, the first condition for the success of the document reading process is insuring confidence between the two parties. Thus document analysis should help in authority evaluation which is based on identifying the names and the roles of the realizers and also studying their behaviours. Our proposal approach is social-based. Besides authority study, it is based on identifying document usage types. A usage type defines how a community of users can access to a document.

Designing a document parser requires to solve complex problems. We argue that authority evaluation requirements and document usage types are important parameters that can reduce this complexity. Thus, the first step of our approach consists in identifying social parameters. Then documents are fragmented regarding their logical structure. We distinguish two levels: the macro-logical level and the micro-logical level. These two levels allow us to design a generic approach that could be applied to documents having different styles and different types.

1 INTRODUCTION

The semantic Web project brought solutions to the limits of the actual Web at the semantic level. In the last few years, Zacklad (2007) introduced the socio-semantic Web with the aim to study the social interactions and how they lead to the creation of explicit and semantically rich knowledge representations. From the system point of view, these interactions can affect many steps in document analysis. As mentioned by Zacklad (2007), the document should be analyzed as a transaction between actors (realizers and beneficiaries) involved in an exchange process.

In this paper, we focus on the problem of analyzing semi-structured documents like scientific papers and books. We believe that reflections presented by Zacklad (Zacklad, 2007) should change our design of approaches and tools of document analysis. Many researchers have and continue to work on analyzing complex documents, like scientific papers and books. Some researchers identify only high level fragments (chapters or sections titles). Others perform detailed analysis but consider only some parts of the document (Connan and Omlin, 2000). Approaches that allow the analysis of the whole document are limited to a document type or style. For example, Rangoni and Belaïd (2006) analyze papers belonging to the same conference.

Despite these works, we think that many questions are still without responses. How to determine elements that should be identified by a document parser? How to define the best granularity level to split a document? Which activity or practice some fragment is useful to?

We believe that a social study can contribute to answer to these questions. Thus, we consider two social dimensions in the document analysis process. The first dimension concerns the “authority notion” considered as a social issue. In fact, the credibility of actors having produced or transmitted a document is one of the relevance criteria. The second dimension is related to the need of personalization based on the user practice.

The remainder of this paper is as follows. Section 2 overviews the problem of document analysis under the social constraints. Our proposed approach is presented in Section 3. Sections 4 and 5 give two applications that illustrate our approach. Finally, Section 6 concludes our paper and suggests some directions for future research.

Bounhas I. and Slimani Y. (2009).
A SOCIAL APPROACH FOR DOCUMENT ANALYSIS.
In *Proceedings of the International Conference on Knowledge Management and Information Sharing*, pages 95-102

2 PROBLEM DISCUSSION

The great number of information suppliers on the Internet and the enormous quantity of the available information caused anxieties towards the information reliability. Information validation, which was done by authors, editors and libraries, is henceforth done by the user. The later is in many situations unable to identify the source of information or to judge its credibility. This difficulty is amplified when many actors having different privileges or prerogatives participate in producing and/or transmitting the information (Zacklad, 2007).

Another change related to the generalisation of the Web is the diversification of the users' needs. From the social point of view, the main phenomenon is the document reading manner or what is called "*document usage*" (Aussenac-Gilles and Condamines, 2004).

Aussenac-Gilles and Condamines (2004) argue that we should model both the text and the usages and that the document usages are not as many as users. Thus, the document usage notion allows viewing the document in a collective perspective which means that a community of users share the same manner of reading a document. Zacklad (2007) confirms this fact by arguing that the actual Web caused the multiplication of document centred collective practices. That's why we consider document usage as a social phenomenon. In fact, user's needs and actions on information (or documents) are related to his/her membership in a community of practice. According to Wenger (1998), a community of practice is "*a group of professionals, informally bound to one another through exposure to a common class of problems, common pursuit of solutions, and thereby themselves embodying a store of knowledge*". According to his membership in a community, a user is interested in some fragments of a document but not in others.

In addition, document analysis should reply to the needs of mediators (such as terminologists) whose task is to analyze its content to produce terminological or ontological resources used to support the research information process and facilitate the end user work.

Thus, we can distinguish three types of actors involved in the documentarisation process:

- Actors participating in document production or transmission,
- Beneficiaries regrouped into communities,
- Mediators: many practices related to these actors such as summarisation, key-word extraction and name entity study.

Our goal is to suggest solutions to the problems of authority evaluation (cf. section 2.1) and the impact of actors practices in document segmentation (cf. section 2.2).

2.1 Authority Evaluation

Authority can be defined as the set of indicators that prove (or can be used to study) the faithfulness of information stakeholders. To study the authority of a web site, we should verify the existence of information such as author names, contact information, copyright texts, and so on.

As mentioned by Zacklad, the identification of such indicators is necessary to document understanding, interpretation and exploitation (Zacklad, 2007).

The authority notion is related to the "*reliability notion*" defined as "*the degree to which the user can trust the information*". Within the web context, many researchers tried to provide methods, metrics and tools for reliability assessment. In fact, it is recognized that authority is the most important dimension of information reliability. Besides on, it is an important relevance criterion (Naumann & Rolker, 2000), (Knight & Burn, 2005) and (Rieh, 2002).

Document authority study is related not only to the identification of intrinsic elements but also to the analysis of extrinsic informations that can help in document evaluation (e.g. biographies of authors). Thus we should identify indicators and then metrics allowing authority evaluation.

2.2 Social Segmentation

For Zacklad, a document is a semiotic production that should have attributes that facilitate practices related to its exploitation. These attributes permit document circulation through communities of interpretation. Zacklad (2007) distinguishes two types of practices and the corresponding attributes types:

- Practices related to the external exploitation of the document when it is stored (and studied) among a collection of documents: For these practices, links between different documents should be identified.
- Practices related to the internal exploitation of the document: to consider these practices, a document should be segmented into coherent parts. Links between fragments should be identified to allow semantic navigation and orientation through the document.

In other hand, we consider that the segmentation of a document depends on the level of granularity chosen. We believe that the *best level* can be defined if we conduct a social study for identifying users' practices. Thus, we argue that the same collection of documents can be modelled and fragmented differently according to the intended usages or the social organization of the users

From the point of view of the socio-semantic Web, a document is a result of a macro communicational transaction composed of many micro-transactions involving many actors. Transactions can be decomposed until a fine level corresponding to the elementary language acts.

According to Zacklad (2007), documents are more fragmented when many actors participate in their production. Besides, micro-transactions are articulated together by elements related to the logical structure such as the document titles and sub-titles, titles numbers and other attributes specifying fragment status or links with other fragments.

We can conclude that the logical structure of a document corresponds to the social process that leaded to its production. That's why the approach that we propose is based on logical structure extraction.

3 THE PROPOSED APPROACH

As conclusion of the above sections, we can claim that reliability insurance is mainly based on authority study which consists in identifying actors who have produced or transmitted the document and their roles. In addition, document segmentation should be based on a social study consisting in identifying users' practices or document usages. This segmentation should be based on the logical structure of the document because this logical structure reflects the communicational transactions that leaded to the production of the document. Finally, document structure study should allow both the internal and external exploitation of the document. Thus, we propose an approach composed of the following steps: social study, document modelling, logical structure extraction and social indexing.

3.1 Social Study

According to the document type and/or the domain, we should, first, study the social process of document production and transmission to identify actors' roles typology in document production and transmission. This study incorporates the identification of any other authority indicators. The second goal of this study is to identify external users' typology and the practices corresponding to each user type (or community). By external users, we mean end users and mediators.

3.2 Document Modelling

The goal of this step is to identify the different document fragment types and the different types of internal and external links. First, we identify fragments types corresponding to internal actors' names and roles. Second, we identify, for each community of end users, fragment and links types needed for their practice. The result of this step is an Xml DTD representing the document model.

3.3 Logical Structure Extraction

The goal of this step is to structure documents according to the document model constructed in the previous step. We fulfil this goal in two steps: macro-logical structure extraction and micro-logical structure extraction. The goal of the first step is to identify high level logical fragments such as chapters or sections. In the second step, we identify the logical entities of each macro-logical fragment by labelling its sub-blocks. The idea behind separating the two levels is to develop a generic and reusable macro-logical analyzer and many micro-logical analyzers each specialized in a analyzing a macro-fragment type.

We should notice that our approach is based on the labelling method widely used in logical structure analysis approaches. At several steps of our analysis process, we define a list of labels that are attributed to blocks using rules. Thus, we inspired from the rule based approach widely used in this filed (Mao, Rosenfeld and Kanungo, 2003).

We use a pre-treatment step where the different types of physical fragments are identified according to the document type and format. We identify three steps: physical analysis, macro-logical structure extraction and micro-logical structure extraction.

3.3.1 Physical Analysis

In this step, the analysed document is restructured according to a physical model shown by figure 1. In fact, we consider that a document is defined by its blocks vector (BV) and a styles vector (SV).

Our model regroups all the blocks types that can be found in a document. Each block type is characterized by several attributes. For example, we compute the number of words of text blocks.

Physical analysis consists, also, in identifying the different styles used in the document. Styles found in the document are added to the SV vector and linked to the corresponding blocks.

D = (BV , SV)
BV = [b1, b2,...,bn]bi ∈ { Text, link, List, Table, Img}
SV = [s1, s2,...,sm]
sj= (Size, Blod, Italic, Uderlined, Color, Align)

Figure 1: The proposed physical model.

3.3.2 Macro-logical Structure Extraction

To recognize the macro-logical structure of a document, we start by identifying macro-logical fragments' titles. To do so, we combine three types of analysis: content analysis, context analysis and style analysis. In fact, we define constraints about the fragment's attributes, the attributes of its predecessor and its successor.

We compute the level of each style by combining three criteria: the style attributes, the style regularity and the style frequency.

When macro-logical titles are identified, we regroup document blocks in a hierarchical manner. We consider that a document is a tree where the top-level elements are chapters or sections according to its type. The leaves of this tree are the physical blocks identified in the Physical analysis step.

This step is performed by using regrouping rules. For example, the following rule regroups a caption with an image into a block labelled "Figure" if the caption is followed by a vertical separator:

```
   Img + Caption_Figure (AfterSep > 0)
==> Figure
```

3.3.3 Micro-logical Structure Extraction

The goal of this step is to identify elementary logical entities required for the different social activities of the intended users. Thus conceiving rules for labelling and regrouping blocks should be based on the social study performed in the first step. For example, according to users' practices, we can consider that a bibliographic reference is an elementary block that should not be segmented. In other cases, we can opt for analyzing each reference to identify its sub-elements. Besides, micro-logical analysis should allow identifying the names and the roles of actors.

Typically a micro-logical analyzer is a semi-structured text parser. In fact, our tool implements an approach based on context free grammars (CFG) which was recently recommended by many researchers. Among the existent solutions, using CFG is the most flexible approach. Besides on, systems based on CFG are able to recognize the text structure especially when it contains complex relationships and constraints between labels (Viola and Narasimhand, 2005).

The grammar used by a micro-logical analyzer is constructed by semi-automatic learning. We made this choice because training data is not available and because we want to help the expert constructing the grammar. Thus, we propose a semi-automatic approach where an expert labels and groups manually the text blocks. The set of labels and rules are inferred from the expert actions to construct the grammar. The expert intervention allows taking into account the social constraints already fixed.

As output is concerned, each analyzer generates an Xml flow that represents the logical structure of the macro-logical fragment. Xml flows generated by micro-logical analyzers are regrouped according to the document model already conceived.

3.4 Social Indexing and Authority Evaluation

By social indexing we do not mean the use of folksonomies. In our sense, social indexing means relating each document or fragment to actors who produced, transmitted or published them. Our goal is to develop tools able to identify names of these actors and their respective roles in existent documents. The document analysis process should also allow identifying any indicator that can used to study the document authority (contact information, copyright texts…etc).

We should remember that authority study is not limited to identifying names and roles of these actors in documents but incorporates the precise identification and the study of the biography of each actor. The goal of this study is to allow authority evaluation by supplying all possible informations about actors. The nature and the size of these informations depend on the domain and/or the intended applications.

The last step consists in analyzing authority indicators. Fundamentally, we want to provide a clear idea about document authority. In better cases can we define metrics for authority evaluation. In the worst cases, should we identify and analyze intrinsic and extrinsic authority indicators and highlight sources of unreliability.

In the following two sections, we present and discuss two examples that illustrate our approach.

4 FIRST APPLICATION: SCIENTIFIC PAPERS

The goal of this project is to create a virtual library of scientific papers. The purpose is to provide adaptable services to the intended users of the library.

4.1 Social Study

Scientific paper production is a typical social phenomenon. Firstly, the scientific progress is realized by accumulating the efforts of researchers around the world each contributing by his effort published in a paper. Secondly, scientific paper production involves many actors having different profiles, responsibilities and roles (the writer, the research head, the team or unit chair…etc). Then, the publication of the paper requires the participation of other types of actors (conference organizers, reviewers, editors, and so on.). When it is published a scientific paper is accessible to a more large community of users what extends its usage.

By analysing the activities of the different actors, we can distinguish two different practices. The first practice is related to papers retrieval, discovery and relevance evaluation. The first criterion for relevance evaluation is the paper's theme. Then, the user would evaluate the paper according to the idea or in general the knowledge it contains. A common reading scenario is based on the paper's logical structure. Thus the user starts by checking the paper's title, then the abstract, after that the introduction and the conclusion and finally the whole content. In other cases, relevance judgment is based on key-words or key-concepts.

To allow documents retrieval and discovery, the virtual library defines two types of research. In the first one, documents are clustered by theme that allows the users to access to the library according to their interests. For the second type of research, documents are indexed by the concepts of a domain ontology which allows finest research. To allow personalised navigation, documents are fragmented and each fragment is related to the concepts of the domain ontology.

The second practice concerns scientific impact evaluation and technological watch. The main element used in scientometry is bibliographic references. The impact of a paper depends on the number of other papers it is cited in. Technological watch means discovering new progresses in a scientific field. The main elements that can be used are the document bibliographic informations, the conclusion and bibliographic references (Dou and al., 1990). Bibliographic informations contains paper's title, authors' names and affiliations what allows discovering new fields and researchers involved in. The conclusion contains, generally, the perspectives that allow identifying new fields of research.

4.2 Document Modelling

From the social study, we can identify the elements that should be involved in a scientific paper model. First, the macro-logical analysis should allow identifying the paper's sections because each section is useful for a different practice. Sections of the paper's body should be identified and organized in hierarchical manner what allows generating the paper's table of contents and thus facilitates navigation in the paper. Because we need to perform conceptual indexing, we should identify more fine logical entities in the body such as paragraphs and captions.

Second, bibliographic informations, references and notes should be analyzed to allow linking different papers and permit bibliometry measurement and technological watch. This analysis allows also identifying actors having participated in the production of the paper (or that are acknowledged) and their respective roles.

The main elements of the obtained document model are shown by Figure 2. It is composed of three parts. The first part called the *front* is composed of the bibliographic information of the paper, the abstract and the key-words. Thus it summarizes the main elements of the paper. We should note that the "*Author*" element regroups author names, roles and affiliations. In the case of papers containing a "*biography*" section, this element regroups also the content of this section. The second element - the *body* - constitutes a detailed view of the paper by regrouping the introduction, the conclusion and the other sections. The last element –the *back*- regroups elements used for scientometry and technological watch such as the bibliographic references and the notes. Each reference or note is decomposed to the finest granularity level to allow authority study and papers linking.

```
<!ELEMENT ScientificPaper (Front, Main, Back?)>
<!ELEMENT Front (title1, title2?,Author+,Abstract*,
Keywords*, Edition?, CopyrightText?)>
<!ELEMENT Main(Introduction?, Section+,
Conclusion?)>
<!ELEMENT Section(title?, (Section | Paragraph | List |
Figure | Table)+)>
<!ATTLIST Section level CDATA >
<!ELEMENT Back (Bibliography?,Notes?,
acknowledgements?)>
<!ELEMENT Bibliography (reference+) >
<!ELEMENT Notes (Note)+>
…
```

Figure 2: Scientific paper model.

4.3 Logical Structure Extraction

Our goal in this step is to prepare the document to be injected into a virtual library where document fragments, themes, concepts and actors are linked together. We performed macro-logical and micro-logical structure extraction according to the above presented approach.

To evaluate our approach, we selected 25 scientific papers in the HTML format from different conferences (and thus having different styles). The prototype was able to analyze a scientific paper in 22.5 seconds as average. At the macro-logical level, 89.09% of the sections were correctly recognized. At the micro-logical level the prototype realized a global success rate equal to 92.85%. These success rates are encouraging if we consider that we used papers having different styles. In fact we recorded better results than those of Rangoni and Belaïd (2006) who recorded a success rate equal to 91.7% for documents sharing the same style.

4.4 Social Indexing and Authority Evaluation

For authority evaluation we consider two criteria. Our prototype permits recognizing bibliographic informations at the finest level of granularity. These informations are linked, for some papers, to biographic informations found in the "*biography*" section. We are planning to construct a database of conferences and journals. At this stage, we plan to integrate information available on the Web and allowing conferences and journals evaluating and ranking. Finally, we plan to develop a tool allowing papers linking based on bibliographic references what will help in evaluating the impact of each paper.

5 SECOND APPLICATION: BOOKS OF ARABIC STORIES

The purpose of this project is to help in Arabic story study. An Arabic story reports historical events or speeches assigned to a person. In the Arabic history, stories were reported among generations by persons called *reporters* or *narrators*. Because these stories report important historical events, Arabic scholars instituted strict rules for story reporting and transmission. Each narrator has an obligation to cite the list of narrators from which he got the story. Thus, the story is preceded by a chain of narrators. Besides on, when a narrator (called the *sheikh*) communicates a story to his follower (the *disciple*), he uses verbs that indicate how he got the story from his predecessor (his *sheikh*) that is called manner of transmission.

Chain of narrators study is a fundamental step that should be performed before studying the story's content. Compared to other types of texts such as scientific papers, authority study gets a greatest importance for Arabic stories. Thus, a story without a chain of narrators (or with an incomplete chain) is rejected whatever is its content.

Stories are actually regrouped in historical books organized by theme. These books were written by specialized scholars who tried to identify reliable stories. Thus, in some cases, the story is followed by comments about the credibility of its narrators and/or a global reliability judgment. The story can also be followed by indications about other versions of the same story. These indications include references about other books where the same story exists. In other cases, these indicators highlight the difference between different versions.

5.1 Social Study

Arabic storytelling is a socio-historical phenomenon characterized by the intervention of many actors having different roles. We can classify these actors as follows:

- Actors participating in the story,
- Actors having reported the story,
- Scholars having collected stories,
- Scholars having judged narrators or stories.

We can identify two types of practices related to arabic study. In one hand, arabic scholars defined strict rules for story acceptance. To be accepted a story should be reported by credible narrators. Narrators are evaluated by specialized scholars based on their behaviours. Twelve classes of credibility are defined. In addition, a story should have a continuous chain of narrators what means that any geographic or temporal gap between two successive narrators is considered as a source of suspect. Thus relations between actors should be studied. At this stage, many types of relations are considered. Besides the kinship and the sheikh-disciple relations, places of residence and dates of birth and death are studied.

On the other hand, an arabic story constitutes a rich source of knowledge. In fact, books of arabic stories support activities of research in arabic history and civilisation.

5.2 Document Modelling

Book of story modelling should allow identifying elements used to story study. As far as authority is concerned two elements should be modelled. In one hand chains of narrators should be identified and parsed to identify names of narrators and manners of transmission. On the other hand, comments about reliability that often follow the story should be recognized and analyzed.

To allow studying stories at the semantic level, the book themes should be recognized. Besides on, story content should be separated from chains and reliability comments to allow conceptual indexing.

As shown by Figure 3, we consider that a historical book is composed of themes and stories. Each story is defined by a chain of narrators, a content and eventually reliability comments and/or versions indicators. As actors are concerned, we developed (by semi-automatic learning) a grammar for full arabic names. This grammar is based on arabic name components described in (Shaalan & Raza, 2007). Besides on, it takes into account social relations between actors (kinship and the sheikh-disciple relation).

```
<!ELEMENT HistoryBook (Theme+) >
<!ELEMENT Theme (Theme | Stroy)+>
<!ELEMENT Story (Chain | Content | Comment |
Version)+>
<!ELEMENT Chain(Actor, Manner)+)>
<!ELEMENT Comment (ActorComment|
StroyComment)>
…
```

Figure 3: Scientific paper model for historical book.

5.3 Logical Structure Extraction

To test our approach we composed a base of 1600 stories selected from four different books. We started by identifying the themes of each book to recognize its macro-logical structure. At this step, we did not record any error. Then, we developed a grammar for parsing an Arabic story. The grammar takes into account actor's names, chains, versions indicators and comments. We evaluated the success rate of our micro-logical parser for four elements as shown by Table 1.

Table 1: Experimentation results for the story parser.

	Success rate
Actors	97.24%
Chains	95.66%
Versions indicators	93.01%
Reliability comments	85.51%

Our results in recognizing arabic person names are better than the results reported by other researchers. For example, the success rate of Zitouni and al. (2005) is equal to 70.2%, Abuleil (2004) reported 83% and Shaalan and Raza (2007) achieved 89%. Being the first to invoke the problem of Arabic story parsing, we could not make a comparison for the remaining elements.

5.4 Social Indexing and Authority Evaluation

Having realized encouraging success rates in story parsing, we are actually working in integrating an existent database of narrators. This database contains exhaustive informations about the narrators of the main arabic historical books. It includes classes of credibility attributed by different scholars to each narrator, information about places and dates of birth and death and information about social relations. Our task is to link names of narrators found in stories to the names of the database. It is a difficult task because an arabic person can have many names and many persons can share the same name(s). Besides, thousands of persons had reported stories which means that narrators' names are very ambiguous. When narrators are exactly identified, we can use the information available in the database to judge the story's reliability. It will be also possible to link different stories based on the narrators they have in common. Then we can perform version comparison. For example when the same story is reported by two narrators, we can compare the two versions and discover any existent anomaly.

6 CONCLUSIONS

In this paper, we discussed the problem of document analysis from a social point of view. We followed the guidelines of Zacklad and other researchers which argue that authority study is one of the main relevance criteria (Zacklad, 2007) and (Rieh, 2002). We were also based on the document usage type notion introduced by Aussenac-Gilles and Condamines (2004). Using these notions, we proposed an approach to extract the logical structure of documents. The first step in this approach is the social study that allows identifying the actors involved in the documentarisation process, document usages and finally fragments and links types to be identified for each document usage. This will lead, in a second step, to a document model that regroups all the required elements. The third step consists in structuring the documents according to this document model. In this step, we distinguished two levels of logical structure: the macro-logical level and the micro-logical level. This distinction allowed us developing reusable and interoperable components that can be used in different document types and styles.

To illustrate our approach, we presented two projects we are actually working on. We used documents in different domains, different types (papers and books) and different languages (Arabic and French).

As future research, we plan to study knowledge extraction and mapping. It includes conceptual indexing and knowledge visualisation. Our goal is to structure a document collection as a map linking themes, fragments, concepts and actors. We think of an intelligent map that guides the user navigation according to his practice represented in his profile. This navigation is also guided by the authority evaluation process.

REFERENCES

Zacklad, M., 2007. Processus de documentarisation dans les Documents pour l'Action (DopA). *Babel - edit -, Le numérique: impact sur le cycle de vie du document*, ENSSIB.

Naumann, F. & Rolker, C., 2000. Assessment Methods for Information Quality Criteria, In. *International Conference on Information Quality (IQ),* Cambridge, MA..

Knight, S. & Burn, J., 2005. Developing a Framework for Assessing Information Quality on the World Wide Web, *Informing Science Journal*, vol. 8, pp. 59-73.

Rieh, S. Y., 2002. Judgment of Information Quality and Cognitive Authority in the Web, *Journal of the American Society for Information Science and Technology*, vol. 53, no. 2, pp. 145-161.

Shaalan, K. & Raza, H., 2007. Person Name Entity Recognition for Arabic. In *ACL'07. Workshop on Computational Approaches to Semitic Languages*, Prague, Czech Republic, pp.17-24.

Viola, P. & Narasimhand, M., 2005. Learning to Extract Information from Semi-structured Text using a Discriminative Context Free Grammar. In *28th annual international ACM SIGIR conference on Research and development in information retrieval*, Salvador, Bahia, Brazil, pp. 330-337.

Rangoni, Y. & Belaïd, A., 2006. Document Logical Structure Analysis Based on Perceptive Cycles. *7th IAPR Workshop on Document Analysis Systems - DAS 2006*, Springer Verlag (Ed.), pp. 117-128.

Dou, H. & Hassanaly, P., Quoniam, L.; La Tela A., 1990. Technological watch and information: on bibliometric analysis in information services, *Documentaliste*, vol. 27, no. 3, pp. 132-141

Connan, J. & Omlin., C. W., 2000. .*Bibliography Extraction with Hidden Markov Models*, technical report US-CS-TR-00-6, computer science department, University of Stellenbosch.

Aussenac-Gilles, N. & Condamines, A., 2004. Documents électroniques et constitution de ressources terminologiques ou ontologiques, *Information-Interaction-Intelligence*, vol. 4, no. 1 pp. 75-94,

Wenger, E., 1998. *Communities of Practice: Learning, Meaning and Identity*, Cambridge University Press.

Zitouni, I., Sorensen, J., Luo, X. & Florian R., 2005. The Impact of Morphological Stemming on Arabic Mention Detection and Coreference Resolution, In *ACL'05, workshop on Computational Approaches to Semitic Languages, 43rd Annual Meeting of the Association of Computational Linguistics*. Ann Arbor, Michigan, USA, pp. 63-70.

Abuleil, S., 2004. Extracting Names from Arabic Text for Question-Answering Systems, In *RIAO'2004, Coupling approaches, coupling media and coupling languages for information retrieval*, Avignon, France. pp. 638- 647.