FUZZY SEMANTIC MATCHING IN (SEMI-)STRUCTURED XML DOCUMENTS

Indexation of Noisy Documents

Arnaud Renard, Sylvie Calabretto and Béatrice Rumpler
Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France

Keywords: Information retrieval, (semi-)Structured documents, XML, Fuzzy semantic matching, Semantic resource, Thesaurus, Ontology, Error correction, OCR.

Abstract: Nowadays, semantics is one of the greatest challenges in IR systems evolution, as well as when it comes to (semi-)structured IR systems which are considered here. Usually, this challenge needs an additional external semantic resource related to the documents collection. In order to compare concepts and from a wider point of view to work with semantic resources, it is necessary to have semantic similarity measures. Similarity measures assume that concepts related to the terms have been identified without ambiguity. Therefore, misspelled terms interfere in term to concept matching process. So, existing semantic aware (semi-)structured IR systems lay on basic concept identification but don't care about terms spelling uncertainty. We choose to deal with this last aspect and we suggest a way to detect and correct misspelled terms through a fuzzy semantic weighting formula which can be integrated in an IR system. In order to evaluate expected gains, we have developed a prototype which first results on small datasets seem interesting.

1 INTRODUCTION

Today's society is evolving and relies on more tools and practices related to information technologies. This is mostly due to the evolution of communication infrastructures. Indeed, the difficulty no longer lies in information availability but rather in access to relevant information according to the user. In order to help in information management, the Web is growing according to two tendencies.

On one side, the first one deals with the larger availability of more structured data. That means that large amounts of data which were formerly stored in flat textual files are now frequently stored in (semi-)structured XML based files. That is the reason why we choose to deal this kind of documents.

On the other side, the second one brings semantic aware techniques in order to achieve better machine level understanding of those data. Semantics consists in the study of words meaning and their relationships like: *homonymy*, *synonymy*, *antonymy*, *hyperonymy*, *hyponymy*. The use of semantics in IR systems can be an efficient way to solve data heterogeneity problems: both in terms of content and data structure representation (documents follow neither the same DTD nor the same XML schema). Most of the time, heterogeneity is due to the lack of a common consensus between information sources, which results in global end users incapacity to have a whole knowledge of documents content and structure in a given collection. As a consequence, semantics can be considered as a key factor in search engine improvement. This observation can be made for both domain specific as well as public at large search engines (like Google, Yahoo …). Indeed, there are an increasing amount of attempts to take semantic into accounts and recently Google integrated some kind of semantics to fill the semantic gap between user real needs and what he has typed as a query. In the same way Microsoft launched recently a new semantic search engine known as "Bing".

It is commonly accepted that the use of semantic resources like ontologies, thesauri and taxonomies of concepts improve IR systems performances (Rosso, 2004). Thus, to use a semantic resource, it is necessary to perform matching between terms of documents and concepts instances in a semantic resource. Some systems already try to achieve (semi-)structured IR by using semantic resources but they are still few. Our goal is to improve results by

Renard A., Calabretto S. and Rumpler B.
FUZZY SEMANTIC MATCHING IN (SEMI-)STRUCTURED XML DOCUMENTS - Indexation of Noisy Documents.
DOI: 10.5220/0002807502530260
In *Proceedings of the 6th International Conference on Web Information Systems and Technology (WEBIST 2010)*, page
ISBN: 978-989-674-025-2

making a fuzzy semantic matching to take into account common mistakes in indexed documents such as typos or wrong words spelling. In fact, none of semantic aware IR systems currently take into account these anomalies.

The article is structured as follows: we present in section 2 some semantic resources, similarity measures, different approaches proposed in the literature about (semi-)structured IR which consider the semantic aspect, and some error correction systems proposition. After, we present our proposal in order to improve semantic indexing of structured documents in section 3. Then, we discuss briefly about prototype design in and the evaluation process in section 4, to finish with evaluation results we obtained in section 5. Finally, we conclude and debate about future works in section 6.

2 RELATED WORKS

A common characteristic of semantic aware IR systems is the necessity of external semantic resources as well as similarity measures allowing for comparisons between concepts. It leads Bellia (Bellia, 2008) to define the notion of semantic framework, which relies on two complementary concepts: an external semantic resource and a model for measuring similarity between concepts.

2.1 Semantic Resources

Semantic resources can be split in two categories according to the range of knowledge they represent: domain specific resources, and general resources. Given the nature of documents collections which are as we indicated before very heterogeneous, only general resources are considered here. Indeed, domain specific semantic resources do not cover a sufficiently broad area and would provide fine grained but fragmentary knowledge about collections. We plan to use general semantic resources: thesaurus like Wordnet, ontologies like YAGO (Suchanek, 2007) which is a large and extensible ontology built on top of Wikipedia and Wordnet, or DBpedia which is resulting from a community effort to extract structured data from Wikipedia (Auer, 2007). DBpedia "uses cases" indicate that it can be used as a very large multilingual and multidomain ontology. DBpedia has the advantage of covering many domains and containing many instances. Moreover, it represents real community agreement and "automatically" evolves as Wikipedia changes. Kobilarov (Kobilarov, 2009) works about interconnection of many domain specific BBC sites by using DBpedia resources seem to be promising. However, there seems to be a lack of semantic similarity measures available on DBpedia data, which makes it difficult to use. As we can see in the next section, semantic similarity measures are very useful to use semantic resources.

2.2 Semantic Similarity Measures

Similarity measures are required to be able to evaluate the semantic similarity of concepts included in a semantic resource such as a thesaurus or ontology. These measures provide estimations about strength of relations between concepts (which queries terms and documents terms are related). It is particularly useful in the semantic disambiguation process, in terms weighting process and when querying by concepts. An almost complete survey of disambiguation may be found in (Navigli, 2009).

Two types of semantic similarity measures can be distinguished. The first type is based on the structure of the semantic resource and counts the number of arcs between two concepts. In contrast, the second type of measures is based on the information content. Information content reflects the relevance of a concept in the collection according to its frequency in the whole collection and the frequency of occurrence of concepts it subsumes. However, Zargayouna (Zargayouna, 2005) showed that the first type of measure could be as efficient as the second one. Moreover, the second type of measures requires a learning phase dependent on the quality of the learning collection. So it is more difficult to carry out (especially because of the difficulty to find a suitable collection for the learning phase). Examples in this area, are Resnik (Resnik, 1995) works who brought the information content, as well as those of Jiang (Jiang, 1997) and Lin (Lin, 1998) using a mixed approach, and more recently of Formica (Formica, 2009). In this work, we will only discuss the first type of measurement.

Rada (Rada, 1989) suggested that the similarity in a semantic network can be calculated by relying on links expressing taxonomic *hypernym/hyponym* relationships, and more specifically of "is-a" type. Then, the semantic similarity can be measured in taxonomy by calculating the distance between concepts by following the shortest path between them. It is mentioned in this article that this method is valid for all hierarchical links ("is-a", "sort-of", "part-of" ...), but it may be modified to take into account other types of relationships.

Wu and Palmer developed in (Wu, 1994) a measure of similarity between concepts for machine translation. Their method is defined according to the distance of two concepts with their smallest common ancestor (the smallest concept that subsumes both of them), and with the root of the hierarchy. The following formula allows computing of similarity between two concepts C_1 and C_2:

$$Sim_{WP}(C_1, C_2) = \frac{2 * depth(C)}{dist(C, C_1) + dist(C, C_2) + 2 * depth(C)} \quad (1)$$

Where, C is the smallest common ancestor of C_1 and C_2 (according to the number of arcs between them), $depth(C)$ is the number of arcs between C and the root, and $dist(C, C_i)$ the number of arcs between C_i and C.

In Zargayouna (Zargayouna, 2005), the proposed similarity measure is based on Wu-Palmer's (Wu, 1994). The "father-son" relationship is privileged over other neighborhood links. To achieve that, Wu-Palmer's measure needs to be modified, because in some cases it penalizes the son of a concept compared to its brothers. Adaptation of the measure is made thanks to the specialization degree function of a concept ($spec$) which represents its distance from the anti-root. This helps to penalize concepts which are not of the same lineage.

$$Sim_{ZS}(C_1, C_2) = \frac{2 * depth(C)}{dist(C, C_1) + dist(C, C_2) + 2 * depth(C) + spec(C_1, C_2)} \quad (2)$$

$$spec(C_1, C_2) = depth_b(C) * dist(C, C_1) * dist(C, C_2) \quad (3)$$

Where, $depth_b(C)$ is the maximum number of arcs between the lowest common ancestor and anti-root "virtual" concept: $\perp$.

In Torjmen (Torjmen, 2008) works on multimedia structure based IR, they assume that the structure of an XML document can be assimilated to ontology. Consequently, they proposed a new refinement of Wu-Palmer (Wu, 1994) and Zargayouna (Zargayouna, 2005) measures applicable directly on documents structure.

Various works designed to manage semantics in IR systems require the use of tools and resources we have introduced. However, most approaches take only the semantic of documents textual content into account and not the semantic of their structure but some IR systems tend to take semantic into account in both content and structure of documents. The XXL system is the first one which incorporated ontology in the indexing process.

2.3 (Semi-)Structured Semantic IR Systems

The XXL query language system allows querying XML documents with syntax similar to SQL. Indeed, it is based on XML-QL and XQuery query languages and adds a semantic similarity operator noted "~". This operator allows expressing constraints of semantic similarity on elements and on their textual content. Query evaluation is based on similarity calculations in ontology as well as terms weighting techniques. The XXL search engine architecture is based on 3 index structures (Schenkel, 2005): element path index, element content index, ontological index. This approach, which consists in semantic indexing by ontology, seems to be interesting.

Van Zwol studies on XSee IR system (Van Zwol, 2007) are interesting because it confirmed that semantic improves the performance of structured IR systems.

Zargayouna (Zargayouna, 2004-2005) works on semantic indexing led to SemIndex prototype (dedicated to the semantic indexing) and SemIR (dedicated to the retrieval). In this system, the semantic dimension is taken into account at both terms and structure levels. The previously defined similarity measure is used for terms sense disambiguation. This is performed favoring the meaning attached to the concept that maximizes the density of the semantic network. The originality of the approach is primarily in the similarity measure used to enrich terms weighting method.

Mercier-Beigbeder measure (Mercier, 2005) is merged by Bellia (Bellia, 2008) with a previous version of Zargayouna's works (Zargayouna, 2004) to take semantic into account. This measure is then enriched to consider XML formalism and latent similarity links between documents.

Other semantic aware structured IR systems may be cited such as CXLEngine (Taha, 2008), which is derived from previous works that led to OOXSearch.

Nevertheless, neither system takes terms uncertainty into account during the indexing process.

2.4 Documents Error Management Mechanisms

Terms uncertainty Errors may have several sources. They can be caused by bad quality documents which results in wrong characters recognition by OCR. Distribution of this kind of errors across documents is somewhat unpredictable a priori. Errors can be caused by human errors in particular when those are

dyslexics, or when they come from foreign countries and learn a new language, or when they write documents on portable devices … Damerau in (Damerau, 1964) established a list of different kind of resulting errors.

According to (Pedler, 2007) two error types can be distinguished: non-words errors which can be easily detected thanks to a dictionary, and real-words errors which are harder to detect while they represent real existing words. Indeed, to be able to detect the second type of errors, the spellchecker must be able to understand (thanks to semantics) the context in which the syntaxically but not semantically correct term is misused.

Error correction problem has been challenged in the Text Retrieval Conference (TREC-5 Confusion Track). Three versions of a collection of more than 55000 documents containing respectively error rates of 0%, 5%, and 20% have been used to run different approaches in the management of those errors for information retrieval systems. A paper which describes this track (Kantor, 2000) indicates the different methods followed by five of the participants. It shows a drop in performances of every IR systems in presence of corrupted documents containing errors. Three of them used query expansion with altered terms and two of them tried to correct documents content. Comparison of these methods indicates that the second approach seems to offer better results and constitute a good starting point. Introduction of semantics in these error correcting systems could be a way to achieve better results. This is the in which our proposal evolves.

3 PROPOSAL: FUZZY SEMANTIC WEIGHTING

During our study of related works, we could identify that Zargayouna's (Zargayouna, 2004-2005) weighting method introduces good concepts for semantic (semi-)structured IR so that we extends (Zargayouna, 2004) semantic weighting formula. The objective is to eliminate mistakes and typos in content by making a fuzzy term matching.

3.1 Terms Semantic Weighting

In Zargayouna (Zargayouna, 2004), the semantic weight $SemW(t,b,d)$ of a term t in a tag b of a document d in the semantic vector corresponds to the sum of its weight and semantically close terms TFITDF weights.

$$SemW(t,b,d) = TFITDF(t,b,d) + \frac{\sum_{i=1}^{n} Sim_{ZS}(t,t_i) * TFITDF(t_i,b,d)}{n} \quad (4)$$

However, TFITDF is better suited for structured XML documents than for (semi-)structured XML documents as it considers specific tags models. Thus, in $SemW_{mod}(t,b,d)$, TFITDF is replaced with standard TFIEFIDF weighting formula.

3.2 Terms Fuzzy Matching

Our idea is to enrich the semantic weighting formula proposed in (Zargayouna, 2004) by taking into account errors in terms spelling leading to uncertainty in written terms. To manage this purpose we were inspired by Tambellini's works on uncertain data management (Tambellini, 2007).

Since we rely on a lexicalized semantic resource where concepts are represented by terms, we believe it may be interesting to perform a fuzzy matching between documents terms and terms reflecting concepts in the lexicalized semantic resource.

According to (Tambellini, 2007), two terms t_1 and t_2 can be paired according to: their concordance i.e. their relative positioning that we note $Conc(t_1,t_2)$, and their intersection i.e. common areas between two terms that we note $Inter(t_1,t_2)$.

Table 1: Adapted Allen's spatial relations.

Relations scheme	Signification	Notation
x y	x *starts* y	*Conc(x,y) = starts*
x y	x *during* y	*Conc(x,y) = during*
x y	x *finishes* y	*Conc(x,y) = finishes*
x y	x *overlaps* y	*Conc(x,y) = overlaps*
x y	x *equals* y	*Conc(x,y) = equals*
x y	x *not_equals* y	*Conc(x,y) = not_equals*

The concordance value noted $ValConc(t_1,t_2)$ is determined according to terms characterization. It depends on spatial relationships derived from Allen's relations (Allen, 1983-1991): "$starts$", "$during$", "$finishes$", "$overlaps$", "$equals$" and "not_equals". Each characterization is then associated with a value α_i.

$$ValConc(t_1,t_2) = \begin{cases} \alpha_1 = 0.8, & if\ Conc(t_1,t_2) = starts \\ \alpha_2 = 0.6, & if\ Conc(t_1,t_2) = during \\ \alpha_3 = 0.8, & if\ Conc(t_1,t_2) = finishes \\ \alpha_4 = 0.2, & if\ Conc(t_1,t_2) = overlaps \\ \alpha_5 = 1, & if\ Conc(t_1,t_2) = equals \\ \alpha_6 = 0, & if\ Conc(t_1,t_2) = not\_equals \end{cases} \quad (5)$$

It should be noted that in (Tambellini, 2007) these values seem to be determined empirically.

We respectively note terms common areas of t_1 and t_2, st_1 and st_2 (cf. Table 1).

The intersection value $ValInter(t_1, t_2)$ is highest i.e. 1 if terms common areas are equals i.e. $st_1 = st_2$ and otherwise its value is $ValInter(st_1, st_2)$.

$$ValInter(t_1, t_2) = \begin{cases} 1, & if\ st_1 = st_2 \\ 0 \leq ValInter(st_1, st_2) < 1, & else \end{cases} \quad (6)$$

The problem of uncertainty is present in many areas including systems which determine if two words are phonetically identical (like *Soundex* algorithm and its derivatives: *Metaphone* …). Spelling correction systems rely on the problem of data uncertainty in the manner they try to compare two words according to their common letters. This kind of algorithm is used to determine $ValInter(st_1, st_2)$. Indeed, terms common areas are phonetically encoded and we note them respectively cst_1 and cst_2:

$$ValInter(st_1, st_2) = 0.75 * \left(1 - \frac{distHamming(cst_1, cst_2)}{\max(length(cst_1), length(cst_2))}\right) \quad (7)$$

The proximity between encodings is computed using a *normalized Hamming distance* and then leveraged with a factor of 0.75 which reflect intersection uncertainty relative to the phonetic encoding. Thus, the matching value $ValApp(t_1, t_2)$ can be defined from $ValConc(t_1, t_2)$ and $ValInter(t_1, t_2)$:

$$ValApp(t_1, t_2) = ValConc(t_1, t_2) * ValInter(t_1, t_2) \quad (8)$$

A term t_1 in a document may be considered to be $present$ in the semantic resource RS if there is a term t_2 in the semantic resource and $Conc(t_1, t_2) = equals$ among concordance relations defined above, and if $ValInter(t_1, t_2) = 1$. In the same way if $Conc(t_1, t_2) = not\_equals$ then it is $missing$ from the semantic resource.

Therefore, it is possible to define an approximation of each term in the documents collection as: all concepts instances of the ontology, which are neither $missing$ nor $present$:

$$\sim t = \{t_{RS} \in C_{RS} | t \approx t_{RS}\} \quad (9)$$

Where, $\sim t$ is the set of terms t_{RS} representing close concepts C_{RS} in the semantic resource RS to a document term t.

3.3 Misleading Terms Detection

It is evident that a fuzzy semantic weighting could introduce noise if applied on correct (not misspelled) terms. Therefore, it is needed to detect off-board terms which can be considered as being misleading terms first. We propose to use Semantic frequency to achieve this by calculating the frequency of the term and that of all semantically close terms:

$$FreqSem(t, b, d) = TF(t, b, d) + \frac{\sum_{i=1}^{n} Sim_{ZS}(t, t_i) * TF(t_i, b, d)}{n} \quad (10)$$

Indeed, if a term t is out of context, its semantic frequency will probably be very low as it will be isolated. Thus, terms whose semantic frequency is below a threshold can be considered as misleading terms. A $\boldsymbol{Context\ Presence\ Indicator}$ function tries to determine if a term t is a wrong term, or not:

$$CPI(t, b, d) = \begin{cases} 1, & if\ FreqSem(t, b, d) > threshold \\ 0, & else \end{cases} \quad (11)$$

The threshold estimation has been experimentally determined and fixed at: $\frac{1}{n} + 0.4 * \frac{1}{n}$, where n is the number of terms in the considered element b. In order to adapt the threshold to different profile of elements textual content (for example when there is not a prevailing thematic in the element), we plan to use an outlier identification data-mining algorithm. The drawback of our error detection method is that it can't detect misleading terms when they are alone in an element. Indeed in that case there is no context in the elements that is why surrounding elements should be used instead.

3.4 Terms Fuzzy Semantic Weighting

Misleading terms detected have to be corrected with best possible substitutes. Our proposition can be considered as fuzzy (imprecise and uncertain) the replacing term selected $t_j \in \sim t$ (among approximations of term t) is the one which seems the most relevant in the context. For this, we select the term t_j whose semantic frequency penalized by its matching value with the term t obtains the highest score:

$$\left\{t_j \in \sim t \middle| \max_{1 \leq j \leq n'} \left(ValApp(t, t_j) * FreqSem(t_j, b, d)\right)\right\} \quad (12)$$

To confer more fuzziness to our proposition we could have considered building a vector of best replacing terms instead of choosing the best one according to our criteria. The new occurrence of the selected replacing term can then be weighted: from nothing if it is not present elsewhere in the element

or its occurrence can be added to the semantic weight of this term if it exists already. Obviously, the matching value is used to weight the significance of the selected term owing to term matching uncertainty. Our terms weighting formula derived from (Zargayouna, 2004) is:

$$SemW_{fuzzy}(t_j,b,d) = \begin{cases} ValApp(t,t_j) * SemW_{mod}(t_j,b,d), & if\ t_j \notin b \\ ValApp(t,t_j) * SemW_{mod,corrected}(t_j,b,d), & else \end{cases} \quad (13)$$

Where $SemW_{mod,corrected}(t_j,b,d)$ is the same formula as $SemW_{mod}(t_j,b,d)$ except the fact that its TFIEFIDF factor is updated to take new possible term occurrence into account.

$$TF_{updated} = \frac{n + ValApp(t,t_j)}{N} \quad (14)$$

$$IEF_{updated} = log\frac{|E|}{|e: t \in e| + ValApp(t,t_j)} \quad (15)$$

$$IDF_{updated} = log\frac{|D|}{|d: t \in d| + ValApp(t,t_j)} \quad (16)$$

We have presented terms fuzzy semantic weighting formula for terms belonging to a document which has been integrated within a semantic aware (semi-)structured IR system to be evaluated.

4 PROTOTYPE

The prototype we have developed to validate our weighting formula should have multiple index structures to access the collection through the structure, the content, and especially the concepts. In addition, it should create a Soundex index of terms in the lexicalized semantic resource in order to make fast comparisons of documents and semantic resources terms phonetic forms. During prototype development phase, we performed a survey of existing libraries and platforms dedicated to IR with the objective to allow the prototype to scale-up on large documents collections. The following tools have been considered: Zettair, Lemur Toolkit, Dragon Toolkit, Terrier, Lucene, GATE. Although none of evaluated systems responds to semantics and (semi-)structured documents constraints, GATE platform has been considered because of its high level of modularity. Thus, it provides many useful tools and libraries to improve prototype development speed, so some of them were used in our prototype (cf. Figure 1). Index persistence problem has been managed through Java Persistence API and a MySQL/InnoDB relational database.

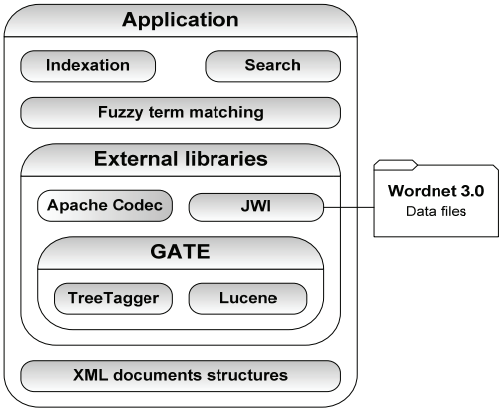

Figure 1: Global application architecture.

5 EVALUATION

The developed prototype allowed us to study the behavior of our proposal against a collection of documents according to the kind of anomalies we wish to correct.

Table 2: Terms distribution per document and element.

Markup	Doc1	Doc2	Doc3	Doc4
name	Anne, Frank	concert	supermarket, grocery, store	movie, theater
sect/title	introduction	introduction	introduction	introduction
sect/par	book, writings, ***dairy***, girl, family, diary, journal	concert, music, musician, recital, ensemble, orchestra, choir, band, show, tour	food, merchandise, meat, produce, dairy, pharmacy, pet, medicine, clothes	movie (x2), theater (x2), theatre (x2), picture, film, cinema, motion, picture, ticket, projector, screen, auditorium

Table 2 is a representation of textual elements of a collection of four documents. Only items with text content are represented as they would result after selection and stemming of words achieved through a morphosyntactic analyzer.

We have deliberately introduced an error in « Doc1/sect/par » element to highlight the interest of our proposal. Indeed, the term “diary” has been replaced by the word “dairy”, shown in red bold italic in Table 2.

5.1 Terms Semantic Frequency

In order to identify off-board terms requiring fuzzy weighting, we calculate the semantic frequency of

each term. The calculated threshold value is 0.2 and is symbolized by a red horizontal line. Each term which semantic frequency is below that point is considered as being off-board, and consequently as misleading.

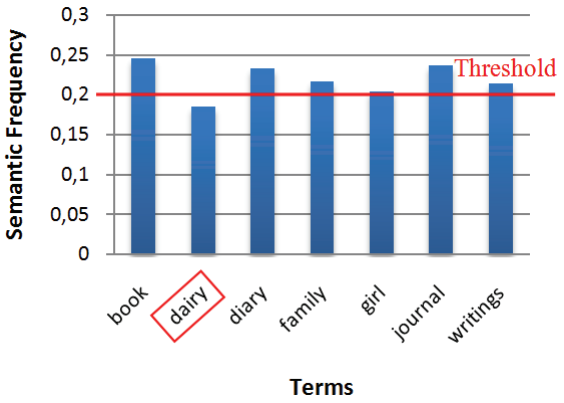

Figure 2: Histogram of terms semantic frequencies.

We can observe on Figure 2 histogram that only one term has a semantic frequency below the threshold. The term "dairy" can be identified as being out of context (it has indeed been introduced as an error on the word "diary"), and therefore the fuzzy semantic weighting formula has to be applied to correct the mistake.

5.2 Terms Fuzzy Semantic Weighting

It is necessary for detected wrong terms to identify the best term in the set of terms approximation. The term of the semantic resource which achieves the best score according to the matching value ($ValApp$) and to its semantic relatedness in the element will be considered as being the best substitute. In the considered case, the best replacing term retrieved for "dairy" is "diary". So the weight of the term "diary" is enriched with "dairy" occurrence. Hence, we have increased the importance of the term "diary" almost as if no mistake occurred on it. Its importance is still lowered due to the uncertainty in terms fuzzy matching.

We can observe on

Figure 3 histogram that none of the first two weighting schemes (solid bars and hollow bars) is able to detect the erroneous writing of an occurrence of the term "diary" spelled as "dairy". This is the normal behavior expected from these formulas. However, we note that the third weighting formula affects to the term "diary" a weighing greater than other terms thanks to enrichment (modulo the confidence of the matching between "dairy" and "diary") of the weighting of this term with the occurrence of erroneous term "dairy".

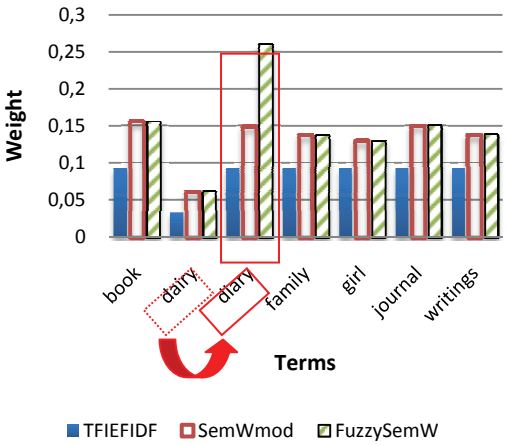

Figure 3: Terms weighting comparison according to the weighting formula.

This is what we want to achieve in order to weight terms beyond errors which can be found in original documents. This can be seen as a semantic corrector which runs during indexation process.

6 CONCLUSIONS AND FURTHER WORKS

In this paper, we have presented a state of the art about useful tools to make semantic aware (semi-)structured IR systems. In particular, semantic similarity measures which allows for concepts comparisons. We then talked about related IR systems and exposed some considerations about error management mechanisms. We finally ended with a proposal for a misleading terms detection method and a fuzzy semantic weighting formula that can be incorporated in an existing system.

The fuzzy matching and weighting formula we propose can be used in conjunction with semantic resources such as Wordnet. An interesting evolution would be to use YAGO or DBpedia instead of Wordnet while they represent much richer resources. Our first evaluations show index quality improvements.

The first short-term development is the implementation of a more scalable prototype allowing us to evaluate error detection/correction and the weighting formula with richer semantic resources on very large datasets like INEX evaluation campaign documents collection.

As indicated before, many other refinements can be considered at different stages. For the misleading term detection, we plan to use data-mining algorithms in order to detect outlier values and avoid the use of empirical thresholds. We plan to include surrounding elements in context definition to help in populating elements context. For the correction phase, we could consider vector of replacing terms instead of choosing the "best" replacing one.

REFERENCES

Allen, J. F., 1983. Maintaining knowledge about temporal intervals. Communications of the ACM, 26(11), 832–843.

Allen, J. F., 1991. Time and time again: The many ways to represent time. International Journal of Intelligent Systems, 6(4), 341–355.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., 2007. Dbpedia: A nucleus for a web of open data. In: LNCS (ed.): Proc. of 6th ISWC, Vol. 4825, Busan, Korea, 722-735.

Bellia, Z., Vincent, N., Kirchner, S., Stamon, G., 2008. Assignation automatique de solutions à des classes de plaintes liées aux ambiances intérieures polluées. Proc. of 8th EGC, Sophia-Antipolis.

Budanitsky, E., Hirst, G., 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. Proc. of 2nd NAACL Workshop on WordNet and other lexical resources.

Formica, A., 2009. Concept similarity by evaluating information contents and feature vectors: a combined approach. Commununications of the ACM 52, 145-149.

Jiang, J.J., Conrath, D.W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. Proc. of International Conference on Research in Computational Linguistics.

Kantor, P., Voorhees, E., 2000. The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text. Information Retrieval, 2(2/3), 165-176.

Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Lee, R., 2009. Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Conections. Proc. of 6th ESWC Semantic Web in Use Track, Crete.

Lin, D., 1998. An Information-Theoretic Definition of Similarity. Proc. of 15th ICML. Morgan Kaufmann Publishers Inc. 296-304.

Mercier, A., Beigbeder, M., 2005. Application de la logique floue à un modèle de recherche d'information basé sur la proximité. Proc. of 12th LFA 2004, 231-237.

Navigli, R., 2009. Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) 41, 1-69.

Pedler, J., 2007. Computer Correction of Real-word spelling Errors in Dyslexic Text. Phd thesis. Birkbeck, London University, 239.

Rada, R., Mili, H., Bicknell, E., Blettner, M., 1989. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics 19, 17-30.

Resnik, P., 1995. Using information content to evaluate semantic similarity in taxonomy. Proc. of 14th IJCAI, 448-453.

Rosso, P., Ferretti, E., Jimenez, D., Vidal, V., 2004. Text categorization and information retrieval using wordnet senses. Proc. of 2nd GWC, Czech Republic, 299-304.

Schenkel, R., Theobald, A., Weikum, G., 2005. Semantic Similarity Search on Semistructured Data with the XXL Search Engine. Information Retrieval 8, 521-545.

Suchanek F., Kasneci G., Weikum G., 2007. Yago – A Core of Semantic knowledge. 16th international World Wide Web conference

Taha, K., Elmasri, R., 2008. CXLEngine: a comprehensive XML loosely structured search engine. Proc. of 11th EDBT workshop on Database technologies for handling XML information on the Web, Vol. 261. ACM, Nantes, France, 37-42.

Tambellini, C., 2007. Un système de recherche d'information adapté aux données incertaines : adaptation du modèle de langue. Phd Thesis. Université Joseph Fourier, Grenoble, 182.

Torjmen, M., Pinel-Sauvagnat, K., Boughanem, M., 2008. Towards a structure-based multimedia retrieval model. Proc. of 1st ACM MIR. ACM, Vancouver, British Columbia, Canada, 350-357.

Van Zwol, R., Van Loosbroek, T., 2007. Effective Use of Semantic Structure in XML Retrieval. In: LNCS (ed.): Proc. of 29th ECIR, Vol. 4425, Rome, Italy, 621.

Wu, Z., Palmer, M., 1994. Verbs semantics and lexical selection. Proc. of 32nd annual meeting of ACL. ACL, Las Cruces, New Mexico, 133-138.

Zargayouna, H., Salotti, S., 2004. Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML. Proc. of IC 2004.

Zargayouna, H., 2005. Indexation sémantique de documents XML. Phd thesis. Université Paris-Sud (Orsay), Paris, 227.