
User Modeling and User-Adapted Interaction **14:** 239–257, 2004.

User Modelling for News Web Sites with Word Sense Based Techniques

BERNARDO MAGNINI and CARLO STRAPPARAVA
ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica, I-38050 Trento, Italy.
e-mail: {magnini, strappa}@itc.it

(Received: 8 October 2002; accepted in final form 5 March 2003)

Abstract. SiteIF is a personal agent for a bilingual news web site that learns user's interests from the requested pages. In this paper we propose to use a word sense based document representation as a starting point to build a model of the user's interests. Documents passed over are processed and relevant senses (disambiguated over WordNet) are extracted and then combined to form a semantic network. A filtering procedure dynamically predicts new documents on the basis of the semantic network.

There are two main advantages of a sense-based approach: first, the model predictions, being based on senses rather than words, are more accurate; second, the model is language independent, allowing navigation in multilingual sites. We report the results of a comparative experiment that has been carried out to give a quantitative estimation of these improvements.

Key words. adaptive hypermedia, content-based user modelling, natural language processing, WORDNET

1. Introduction

SiteIF (Stefani and Strapparava, 1998; Strapparava et al., 2000; Magnini and Strapparva, 2001) is a personal agent for a multilingual news web site, that takes into account the user's browsing by 'watching over the user's shoulder'. It learns user's interests from the requested pages that are analyzed to generate or to update a model of the user. Exploiting this model, the system tries to anticipate which documents in the web site could be interesting for the user.

Adaptive web-based information filtering is an emerging and important research field (Micarelli and Sciarrone, 2004; Waern, 2004). Many systems (e.g., (Lieberman et al., 1999; Minio and Tasso, 1996)) that exploit a user model to propose relevant documents build a representation of the user's interest which takes into account some properties of words in the document, such as their frequency and their co-occurrence. However, assuming that interest is strictly related to the semantic content of the already seen documents, a purely word based user model is often not accurate enough. The issue is even more important in the web world, where documents have to do with many different topics and the chance to misinterpret word senses is a real problem.

In this paper we propose the use of a sense-based document representation to build a model of the user's interests. As the user browses the documents, the system builds

the user model as a semantic network whose nodes represent senses (not just words) of the documents requested by the user. Then, the filtering phase takes advantage of the word senses to retrieve new documents with high semantic relevance with respect to the user model.

The use of senses rather than words implies that the resulting user model is not only more accurate but also independent from the language of the documents browsed. This is particularly important for multilingual web sites, that are becoming very common especially in news sites or in electronic commerce domains.

The sense-based approach adopted for the user model component of the SiteIF system makes use of WORDNET DOMAINS (Artale et al., 1997) (Magnini and Cavaglià, 2000), a multilingual lexical database where English and Italian senses are aligned and where each sense is annotated with *domain labels* (such as MEDICINE, ARCHITECTURE and SPORT). A technique, recently proposed in (Magnini and Strapparava, 2000; Magnini et al., 2002), called Word Domain Disambiguation, has been adopted to disambiguate the word senses that define the user interest model. Although SiteIF focuses on news recommendation, this sense-based approach may be broadly applicable. For example, it may be appropriate also for e-mail, movies/books recommendation filtering and, in general, whenever a text or structured description of content is available.

As for the filtering phase, our approach is supported by experimental evidence (e.g., (Gonzalo et al., 1998a)) showing that a content based match (exploiting word meanings) can significantly improve the accuracy of the retrieval.

The paper also describes an empirical evaluation of a word meaning based versus a traditional word-based user modelling. This experiment shows a substantial improvement in performance with respect to the word based approach.

The paper is organized as follows. Section 2 gives a sketch of the kind of documents the system deals with and describes how WORDNET DOMAINS and the disambiguation algorithms can be exploited to represent the documents in terms of lexical concepts. Section 3 describes how the user model is built, maintained and used to propose new relevant documents to the user. Section 4 provides some notes about system implementation and interface. Section 5 gives an account of the experiment that evaluates and compares a synset-based user model versus a word-based user model. Some final comments about future developments conclude the paper.

2. Sense-Based Document Representation

The SiteIF web site has been built using a news corpus kindly put at our disposal by ADNKRONOS, an important Italian news provider. The corpus consists of about 5000 parallel news (i.e., each news has both an Italian and an English version) partitioned by ADNKRONOS in a number of seven fixed categories: `culture`, `food`, `holidays`, `medicine`, `fashion`, `motors` and `news`. These categories were chosen by ADNKRONOS to partition the corpus and have no explicit connection with the domain labels in WORDNET DOMAINS. The average length of the news is about 265 words. Figure 1 shows an example of parallel (English-Italian) news.

The main working hypothesis underlying our approach to user modelling is that a sense-based analysis of the document can improve the accuracy of the model. There are two crucial questions to address: first, a repository for word senses has to be identified; second, the problem of word sense disambiguation, with respect to the sense repository, has to be solved.

Section 2.1 introduces the sense repository we use. Section 2.2 gives some details about WDD, while Section 2.3 shows how WDD is applied to represent documents in our system.

2.1. SENSE REPOSITORY

As for sense repository we started from WORDNET (version 1.6) (Miller, 1995; Fellbaum, 1998), a large lexical database for English, freely available, which has received a lot of attention within the computational linguistics community. Nouns, verbs, adjectives and adverbs are organized into synonym sets (i.e., *synsets*), each representing one underlying lexical concept. Figure 2 shows a sketch of the lexical matrix underlying WORDNET. In general the mappings among lemmas and meaning are many to many. For example, L_1 and L_2 are synonymous while L_2 is polysemous. The rows M_i in the matrix represent synsets. Polysemy and synonymy are phenomena related to access information in the mental lexicon: a listener/reader who recognizes a form must cope with polysemy a speaker/writer who express a meaning must decide between synonyms.

CULTURE: GIOTTO- PAID BY MONKS TO WRITE ANTI-FRANCISCAN POETRY
Rome,10 Jan. -(Adnkronos)- Giotto was 'paid' to attack a faction of the Franciscans, the Spiritual ones, who opposed church decoration in honour of Poverello di Assisi. This has been revealed in the research of an Italian scholar who is a professor at Yale University, Stefano Ugo Baldassarri, who thinks he has solved the mystery of the only known poetry by the famous Tuscan painter: the Giotto verses have in fact always provoked wonder because they seem to be a criticism of the ideals of St. Francis and all the more so since their author was also the man who painted the famous frescoes of the Basilica at Assisi. ...

CULTURA: GIOTTO- PAGATO DA FRATI PER SCRIVERE POESIA ANTI-FRANCESCANA
Roma, 10 gen. -(Adnkronos)- Giotto fu 'pagato' per attaccare una fazione dei Francescani, quella degli Spirituali, che si opponevano alla decorazione delle chiese in onore del Poverello di Assisi. Lo rivela una ricerca di uno studioso italiano docente alla Yale University, Stefano Ugo Baldassarri, che ritiene di aver svelato il mistero dell'unica poesia conosciuta del celebre pittore toscano: i versi giotteschi, infatti, avevano sempre destato meraviglia perché apparivano come una critica agli ideali di San Francesco, tanto piu' mossa proprio dall'autore dei celebri affreschi della Basilica di Assisi. ...

Figure 1. Sample of parallel news texts.

	Lemmata				
Word Meanings	L_1	L_2	L_3	$\cdots$	L_n
M_1	$E_{1,1}$	$E_{1,2}$			
M_2		$E_{2,2}$			
M_3			$E_{3,3}$		
$\vdots$				$\ddots$	
M_m					$E_{m,n}$

Figure 2. WORDNET lexical matrix.

Synsets are linked by different semantic relations (IS-A, PART-OF, etc...) and organized in hierarchies. The synsets identified in WORDNET derive from a long lexicographic work and many fine-grained sense distinctions are made (there are 99,642 synsets in version 1.6).

The main advantage in using WORDNET is that versions in languages other than English are now available for a number of European languages, including Spanish, German, Basque, Catalan, Dutch, Estonian and Italian. Even if none of these wordnets has the coverage of the English WORDNET, they are suitable to be used in several application scenarios, as it emerged from the contributions presented at the recent Global WordNet Conference (Fellbaum and Vossen, 2002). In particular, in SiteIF we use WORDNET DOMAINS, a multilingual extension of WORDNET 1.6, developed at ITC-irst, based on the assumption that the semantic relations already defined for the original English version may, for the most part, be reused for other languages.[1] From an implementation point of view, each synset has an English and an Italian part, sharing the same semantic organization (e.g., IS-A, PART-OF, etc... relations).

As well in WORDNET DOMAINS each synset has been annotated with at least one domain label, selected from a set of about two hundred labels hierarchically organized (see (Magnini and Cavaglià, 2000) for the annotation methodology and for the evaluation of the resource).

A domain may include synsets of different syntactic categories: for instance MEDICINE groups together senses from Nouns, such as `doctor#1` (i.e., the first sense of the word *doctor*) and `hospital#1`, and from Verbs such as `operate#7`. Second, a domain may include senses from different WORDNET sub-hierarchies (i.e. deriving from different 'unique beginners' or from different 'lexicographer files'). For example, SPORT contains senses such as `athlete#1`, deriving from `life_form#1`, `game_equipment#1`, from `physical_object#1` `sport#1` from `act#2`, and `playing_field#1`, from `location#1`.

[1]This a strong assumption and it may be considered plausible if we limit ourselves to the main indoeuropean languages, among which there is much cultural overlap (Miller 1997 – personal communication).

Finally, domains may group senses of the same word into homogeneous clusters. As we see later, this grouping gives a substantial help in designing the word sense disambiguation algorithm. For the WSD algorithm in the SiteIF system we have considered 41 disjoint domain labels which allow a good level of abstraction without loosing relevant information (i.e., in the experiments we have used SPORT domain label in place of VOLLEY or BASKETBALL, which are both subsumed by SPORT). Note that choosing more general domain labels has the effect of obtaining larger clusters of homogeneous synsets, but the senses, that we use to build document representations, are the synsets with their granularity.

WORDNET DOMAINS is also a multilingual extension of the English WORDNET. The Italian part of WORDNET DOMAINS currently covers about 40,000 lemmas, completely aligned with the English WORDNET (i.e., with correspondences to English senses).

The advantages of a synset-based document representation are that: (i) ambiguous terms in the document are disambiguated, therefore allowing their correct interpretation and consequently a better precision in the user model construction (e.g., if a user is interested in financial news, a document containing the word 'bank' in the context of geography will not be relevant); (ii) synonym words belonging to the same synset can contribute to the user model definition (for example, both 'bank' and 'bank building' bring evidences for financial documents, improving the coverage of the document retrieval); (iii) finally, as we use a multilingual wordnet, synsets will match with synonyms in various languages, allowing the user model to be defined on a multilingual base.

2.2. WORD DOMAIN DISAMBIGUATION

As far as word disambiguation is concerned, we have used Word Domain Disambiguation (WDD), a technique proposed in (Magnini and Strapparava, 2000) based on sense clustering through the annotation of the WORDNET DOMAINS synsets with domain labels. We have addressed the problem starting with the hypothesis that domain information is useful to reduce the complexity of word sense disambiguation. This line is also supported by several works (see for example (Stevenson and Wilks, 2001), (Gonzalo et al., 1998b), (Kilgarriff and Yallop, 2000) and the SENSEVAL initiative) which remark that for many practical purposes (e.g., cross lingual information retrieval) the fine-grained sense distinctions provided by WORDNET are not always necessary.

Word Domain Disambiguation is a variant of Word Sense Disambiguation where the role of *domain information* is exploited. The hypothesis is that domain labels (such as MEDICINE, ARCHITECTURE and SPORT) provide a natural and powerful way to establish semantic relations among word senses, which can be profitably used during the disambiguation process. In particular, domains constitute a fundamental feature of textual coherence, such that word senses occurring in a coherent portion of text tend to maximize domain similarity.

Table I shows an example of how domain labels can provide a natural way to group word senses into homogeneous clusters. The word 'bank' has ten different senses in WORDNET 1.6: three of them (i.e., sense 1, 3 and 6) can be grouped under the ECONOMY domain, while sense 2 and 7 are both belonging to GEOGRAPHY and GEOLOGY.

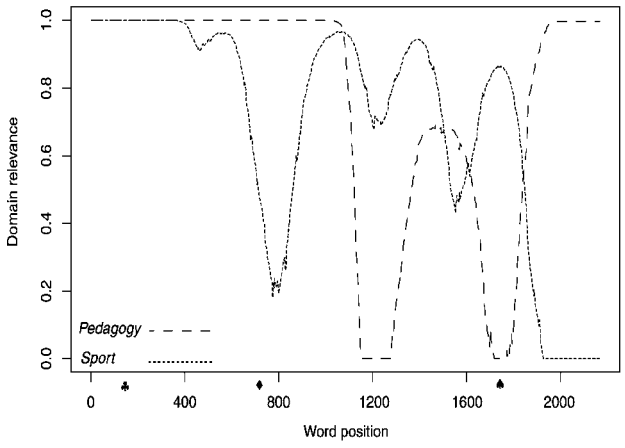
The starting point in the algorithm design was the previous work in word domain disambiguation reported in (Magnini and Strapparava, 2000). The basic idea was that the whole disambiguation process can be profitably decomposed into two tasks: first, choosing a domain label for the target word among those reported in WORDNET DOMAINS; then, select a sense of the target word among those which are compatible with the selected domain. The algorithm was in two steps. Each word in the text is considered and for each domain label allowed by that word a score is given. This score is determined by the frequency of the label among the senses of the word. At the second step each word is reconsidered, and the domain label with the highest score is selected as the result of the disambiguation. In (Magnini and Strapparava, 2000) it is reported that this algorithm reaches a good accuracy in word domain disambiguation, both for Italian and English, on a corpus of parallel news. This result makes WDD appealing for applications where fine-grained sense distinctions are not required, such as document user modelling.

However, one drawback of this algorithm is that, for rather long texts, it does not consider domain variations. The present description, to overcome these problems, considers portions of text (i.e., *contexts*) within which domain relevance is calculated. (Figure 3 shows an example of domain variation detected by the WDD algorithm).

A second direction of work has been the acquisition of domain information from annotated texts (i.e., Semcor, a portion of Brown corpus annotated with WORDNET senses).

Table I. WORDNET senses, domains and occurrences in Semcor for the word 'bank'

Sense	Synset & Gloss	Domains	Semcor occurr.
#1	Depository financial institution, bank, banking concern, banking company (a financial institution...)	ECONOMY	20
#2	Bank (sloping land...)	GEOGRAPHY, GEOLOGY	14
#3	Bank (a supply or stock held in reserve...)	ECONOMY	–
#4	Bank, bank building (a building...)	ARCHITECTURE, ECONOMY	–
#5	Bank (an arrangement of similar objects...)	FACTOTUM	1
#6	Savings bank, coin bank, money box, bank (a container...)	ECONOMY	–
#7	Bank (a long ridge or pile...)	GEOGRAPHY, GEOLOGY	2
#8	Bank (the funds held by a gambling house...)	ECONOMY, PLAY	–
#9	Bank, cant, camber (a slope in the turn of a road...)	ARCHITECTURE	–
#10	Bank (a flight maneuver...)	TRANSPORT	–



♣ ...The Russians are all trained as dancers before they start to study gymnastics. ...

♦ ...If we wait until children are in junior-high or high-school, we will never manage it. ...

♠ ...The backbend is of extreme importance to any form of free gymnastics, and, as with all acrobatics, the sooner begun the better the results. ...

Figure 3. Domain variation in the text *br-e24* from the Semcor corpus.

Results obtained at the SENSEVAL-2 competition (SENSEVAL-2, 2001) on Word Sense Disambiguation confirm that for a significant subset of words, domain information can be used to disambiguate with a very high level of precision. For the three task we participated in, i.e., English All Words (all the words in the documents have to be disambiguated), English Lexical Sample (only sample words have to be disambiguated) and Italian Lexical Sample, no other syntactic or semantic information has been used (e.g., semantic relations in WORDNET) but domain labels. As far as English All Words task is concerned (i.e., the task relevant in the context of SiteIF system) our algorithm obtained .748 and .357 in terms of precision and recall respectively. The precision was the best among all the systems participating to this task. The rather low degree of recall reflects the fact that just few words in a text carry relevant domain information. In fact, the algorithm does not take into account 'generic' senses and disambiguates those terms that are relevant with respect to the document topics.

A promising direction is to take advantage of multilinguality, in particular of the fact that we have parallel bilingual news, to improve the performance of the disambiguation algorithm. In (Magnini and Strapparava, 2000) there are some preliminary results using a *mutual help* disambiguation strategy, which takes advantages of differential polisemy among languages and of the shared senses of parallel bilingual texts.

The following sections presents details of the disambiguation procedures implemented for last version of SiteIF.

2.2.1. *Linguistic Processing*

As for lemmatization and part-of-speech tagging the Tree Tagger, developed at the University of Stuttgart (see (Schmid, 1994)) has been used, both for English and Italian. The WORDNET morphological analyser has been also used in order to resolve

ambiguities and lemmatization mistakes. After this process texts are represented as vectors of triples: word lemma, WORDNET part of speech, position in the text.

2.2.2. *Scoring Domains for a Lemma*

The basic procedure in domain driven disambiguation is a function that, given a lemma L, associates a score to each domain defined for that lemma in WORDNET DOMAINS. Such a score is the relative frequency of the domain in L, computed on the basis of the occurrences of the synsets of L in Semcor. Semcor occurrences for synsets with multiple domain annotations are repeated for each domain, while synsets with 0 occurrences are counted as 0.5. As an example, consider the lemma 'bank' in Table I. The total occurences are 57. Table II shows the domain frequencies for that lemma. For example, the GEOLOGY domain collects contributions from senses 2 and 7, for a total of 16 occurrences in Semcor, which corresponds to a frequency 0.28 (i.e., $fq[D_{Geology}](bank) = 0.28$).

2.2.3. *Domain Vectors*

The data structure that collects domain information is called *Domain Vector* (DV). Intuitively a DV represents the domains that are relevant for a certain lemma (or word sense) in a certain context. We have considered three kinds of DV's: a DV for a lemma L within a context C (DV_L^C), as it is the case of test data; a DV for a synset S of a lemma L within a context C (DV_S^C), as it is the case of training data; a DV for a synset S of a lemma L in WORDNET (DV_S), which is used when no training data are available.

– DV *for a lemma in context* (DV_L^C)

Given a set of domains $D_1 \ldots D_n$, a DV for a lemma L in a position K within a text, represents the relevance of those domains for that lemma, i.e., each component $DV_L[i]$ gives the degree of relevance of the domain D_i for the lemma L. Given a context of $\pm C$ words before and after the lemma L in the position K, each component of the domain vector is defined with the following formula:

$$DV_L^C[i] = \sum_{k=K-C}^{K+C} fq[D_i](L_k) * gauss(k)$$

Table II. Domain frequencies for 'bank'

Wordnet domain	Scoring
ECONOMY	.38
GEOGRAPHY	.28
GEOLOGY	.28
FACTOTUM	.02
ARCHITECTURE	.02
PLAY	.01
TRANSPORT	.01

where *gauss* is the normal distribution centered in the position K. In the current algorithm C is set to 50.

Intuitively, the above formula takes into account the contribution of the lemmas in the context C to the sense of the target lemma L. In addition a DV actually selects a set of relevant domains rather then just one domain. The complexity of the computation for a text T is $length(T) * 2C * |\mathcal{D}| * \varepsilon$ where $length(T)$ is the number of words contained in the text T, $\mathcal{D}$ is the set of considered domains (i.e., in our case the 41 domains), C is the context and ε is a constant to access the domain frequencies for a lemma (which are tabularized).

– DV *for a synset in context* (DV_S^C)

In case a training corpus is available, where lemmas are annotated with the correct sense, Domain Vectors are computed with the formula above. Instead of considering a lemma in a position K within a text we have a sense for that lemma (i.e., a synset). DV_S^C represents a 'typical' vector for a sense S of a lemma L.

– DV *for a synset without context* (DV_S)

When a training corpus is not available (as for the all words task), the simpler way to build a DV for a certain synset is to compute it with respect to WORDNET DOMAINS. Given a synset S, the domain vector DV_S is a vector that has 1's in the position of its domain(s) and 0's otherwise. More accurate DV could be obtained considering contextual information such as the synset gloss. This kind of vectors can be built only one time, so every access to them can be considered constant.

2.2.4. *Comparing Domain Vectors*

To disambiguate a lemma L (i.e., the target lemma) in a text, first its DV_L^C is computed. The next step consists in comparing the DV of the target lemma L with the domain vectors for each sense of L derived either from the training set, when available, or from WORDNET DOMAINS, when training data are not available. The sense vector DV_S which maximize the similarity is selected as the appropriate sense of L in that text. The similarity between two DV's is calculated with the standard dot product: $\mathrm{DV}_1 \cdot \mathrm{DV}_2 = \sum_i \mathrm{DV}_1[i] * \mathrm{DV}_1[i]$. As far as complexity is concerned, note that, given a lemma, we have as many dot products as senses of that lemma. In the corpus news used in the present version of SiteIf, the mean polisemy is about 2.88.

2.3. DOCUMENT REPRESENTATIONS

Each document maintained in the SiteIF site is processed to extract its semantic content. Given that we rely on WORDNET DOMAINS, the final representation consists of a list of synsets relevant for a certain document. The text processing is carried out whenever a new document is inserted in the web site, and includes two basic phases: (i) lemmatization and part-of-speech tagging; (ii) synset identification with WDD.

The lemmatization and part-of-speech tagging of the document constitutes the input for the synset identification phase, which is mainly based on the word domain disambiguation procedure described in Section 2.2. The WDD algorithm, for each word, proposes the domain label and the synsets appropriate for the word context.

As we noted in the previous section, the time efficiency of the WDD algorithm is quite reasonable to produce a complete document representation as soon as the document enter the web site.

As an example, Figure 4 shows a fragment of the Synset Document Representation (SDR) for the document presented in Figure 1. Words are presented with the preferred domain label as well as with the selected synsets. For readability reasons we show the synonyms belonging to each synsets in place of the synset unique identifier used in the actual implementation. In addition, only the English part of the synset is displayed.

Being based on synsets, SDR's are portable through the languages supported by aligned wordnets. This allows to build user models for multilingual sites.

Word lemma	*Domain label*	*Synsets*
faction	Factotum	{faction-2, sect-2} {cabal-1, faction-1, junta-1, junto-1, camarilla-1}
franciscan	Religion	{Gray_Friar-1, Franciscan-1}
church	Religion	{church-1, Christian_church-1, Christianity-2} {church-2, church_building-1} {church_service-1, church-3}
decoration	Factotum	{decoration-3}
honour	Factotum	{award-2, accolade-1, honor-1, honour-2, laurels-1} {honor-3, honour-4}
research	Factotum	{research-1} {inquiry-1, enquiry-2, research-2}
scholar	Pedagogy	{scholar-1, scholarly_person-1, student-2} {learner-1, scholar-2} {scholar-3}
professor	Pedagogy	{professor-1}
mystery	Literature	{mystery-2, mystery_story-1, whodunit-1}
poetry	Literature	{poetry-1, poesy-1, verse-1} {poetry-2}
painter	Art	{painter-1}
verse	Literature	{poetry-1, poesy-1, verse-1} {verse-2, rhyme-2} {verse-3, verse_line-1}
wonder	Factotum	{wonder-2, marvel-1}
criticism	Factotum	{criticism-1, unfavorable_judgment-1}
ideal	Factotum	{ideal-1} {ideal-2}
man	Factotum	{man-1, adult_male-1} {man-3} {man-7} {man-8}
author	Literature	{writer-1, author-1}
fresco	Art	{fresco-1} {fresco-2}
basilica	Religion	{basilica-1}

Figure 4. Synset Document Representation for a fragment of the text shown in Figure 1.

3. Sense-Based User Modelling

There are much work and many approaches addressing the general problem of sense-based document representation/description. Most of these approaches are mainly developed in communities that do not primarily focus on user modelling, and are heavily based on simple phrase statistics and/or machine learning techniques. In particular (supervised) machine learning techniques involve the problem of preparing training sets of documents, so requiring hand-tagging of senses. This extra manual work is simply not feasible in the dynamic context of news web sites, in which the categories and the number of documents are subject to change in an unexpected way.

As seen in the previous section, our word sense disambiguation algorithm can run completely unsupervised, it has a reasonable computational overhead and it is based on a (very) simple notion of meanings (i.e., the synsets) in a 'standard' repository (i.e., WORDNET).

The task of building an appropriate model of the user exploiting (possibly multilingual) documents shares some of the goals with cross-lingual information retrieval (see for example (Grefenstette, 1998) for a survey), but it is less complex in that it does not require a query disambiguation process. In fact in our case the matching is between a fully disambiguated structure, i.e., the user model, and the document contents.[2] Anyway, the idea of indexing documents by means of WORDNET synsets is an emerging tendency, supported both by the availability of multilingual lexical resources with a clear semantic structure, such as WORDNET DOMAINS or EuroWordnet (Vossen, 1998), and by experimental evidences (e.g., (Gonzalo et al., 1998a)), which shows that a word sense based match can improve the accuracy of the retrieval with a set of manually disambiguated texts.

In SiteIF the user model is implemented as a semantic net whose goal is to represent the contextual information derived from the documents. Previous versions of SiteIF were purely word-based, that is the nodes in the net represented the words and the arcs the word co-occurrences. However the resulting user models were fixed to the precise words of the browsed news. One key issue in automating the retrieval of potentially interesting news was to find document representations that are semantically rich and accurate, keeping to a minimal level the participation of the user.

A new version of SiteIF has been realized where the user model is still implemented as a network structure, with the difference that nodes now represent synsets and arcs the co-occurrence of synsets. The working hypothesis is that the model can help to define semantic chains through which the filtering has a better chance to catch documents semantically closer to the topics already touched by the user.

Possibly modelling with synsets or with words will bring to different choices and optimizations in the semantic network representation. However in this paper one purpose is to compare the results of word-based and of synset-based user model,

[2]For an overview of the issue to use word-sense disambiguation techniques in a pure IR environment see, for example, (Krovetz and Croft, 1992; Voorhees, 1993), while for WSD in query expansion for cross language IR useful references are (Ballesteros and Croft, 1997; Hull, 1997).

and then we keep uniform the machinery of the user model data structures and algorithms.

3.1. MODELLING PHASE

SiteIF falls into the category of systems that manage personalized views of information spaces (see (Brusilovsky, 1998)). It provides adaptive recommendations on a closed corpus (i.e., a single web site), but the corpus is highly dynamic, i.e., the news web site can expand and change gradually but in principle with no limit. The system can cope with this, as the overhead to build a document representation in term of synsets is acceptable.

Our sense-based recommendation approach yields a dynamic profile of the user's interest in terms of the words and phrases meanings that differentiate items of interest to the user from other items. While SiteIF focuses on news recommendation, this technique is appropriate also for e-mail, movies/books recommendation filtering and, in general, every time a text or structured description of content is available.

In the modelling phase SiteIF considers the browsed documents during a user navigation session (one user model is maintained for each user). The system uses the document representation of the browsed news. Every synset has a score that is inversely proportional to its frequency over all the news corpus. The score is higher for less frequent synsets, avoiding that very common meanings become too prevailing in the user model. Likewise, in the word-based case we considered a word list document representation, where every word has a score inversely proportional to the word frequency in the news corpus.

The system builds or augments the user model as a semantic net whose nodes are synsets and arcs between nodes are the co-occurrence relation (i.e., co-occuring presence in a document) of two synsets. Synset co-occurence represents the simultaneous presence in a document of those meanings. This is a simplification, however the news are short enough not to be a disadvantage. Weights on nodes are incremented by the score of the synsets, while weights on arcs are the mean of the connected nodes weights.[3] For each browsed news, the weights of the net are periodically reconsidered and possibly lowered, depending on the time passed from the last update. Also no longer useful nodes and arcs may be removed from the net. In this way it is possible to consider changes of the user's interests and to avoid that uninteresting concepts remain in the user model.

Figure 5 sketches the modelling process showing an example of user model augmentation.

[3]As far as the arcs are concerned, an indication of the semantic similarity between the synsets using word sense disambiguation techniques is also present. This is useful to build cohesive chains of synsets in the user model network. (See (Resnik, 1995) for an introduction about word sense disambiguation and semantic similarity issue.) However, we do not take advantage of this information in the evaluation experiment in Section 5.

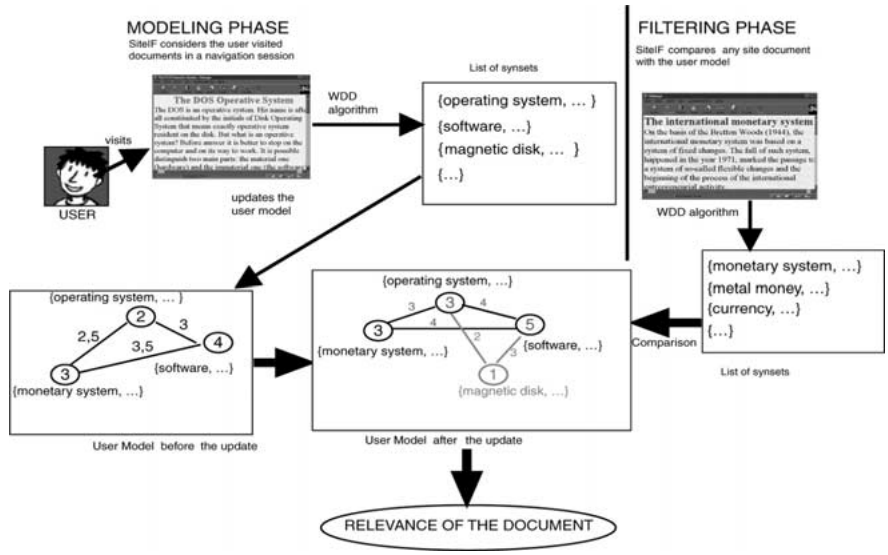


Figure 5. Modelling and Filtering Processes.

3.2. FILTERING PHASE

During the filtering phase the system compares with the user model any document (i.e., the representation of any documents in terms of synsets) in the site. A matching module receives as input the internal representation of a document and the current user model and it produces as output a classification of the document (i.e., whether it is worth or not the user's attention). The relevance of any single document is estimated using the Semantic Network Value Technique (see for details (Stefani and Strapparava, 1998)). The idea behind the SiteIF algorithm consists of checking, for every concept in the representation of the document, whether the context in which it occurs has been already found in previously visited documents (i.e., those already stored in the semantic net). This context is represented by a co-occurrence relationship, i.e., by the couples of terms included in the document which have already co-occurred before in other documents. This information is represented by arcs of the semantic net.

Here below we present the formula used to calculate the relevance of a document using the Semantic Network Value Technique:

$$\text{Relevance}(\text{doc}) = \sum_{i \in \{\text{syns}(\text{doc})\}} w(i) * \text{freq}_{\text{doc}}(i) + \sum_{i,j \in \{\text{syns}(\text{doc})\}} w(i,j) * w(j) * \text{freq}_{\text{doc}}(j)$$

where $w(i)$ is the weight of synset-node i in the UM network, $w(i,j)$ is the weight of the arc between i and j.

See Figure 5 for a summary sketch of the filtering process and Figure 6 for a detailed example of relevance calculation.

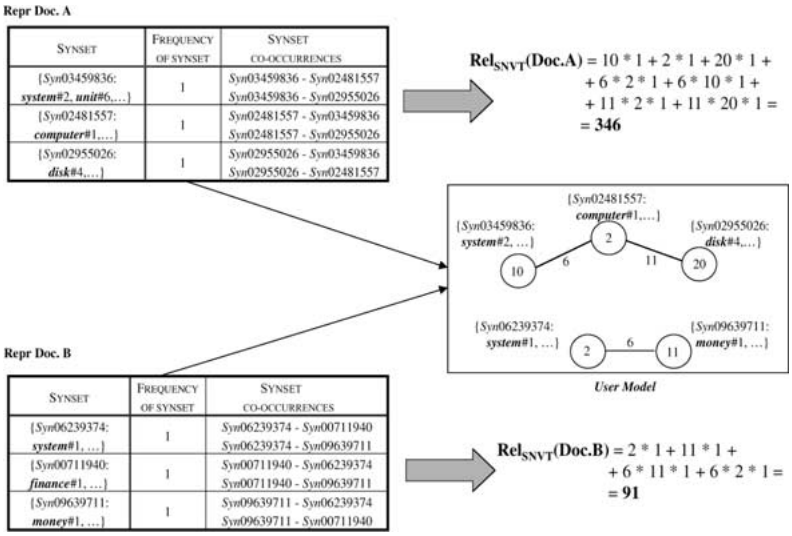


Figure 6. Relevance of two documents according to the user model (each synset has a unique identifier Syn*nnnn*).

4. The SiteIF Prototype

The SiteIF system is entirely implemented in Common Lisp exploiting the CL-HTTP package. CL-HTTP is a full-featured server developed at MIT for the Internet Hypertext Transfer Protocol that comes free and complete with source code.[4] The server has been proven in major production systems and applied in a number of Artificial Intelligence systems.

SiteIF first requires the user to log in (through username and password) and, once verified, it allows the user to enter the web site. Then, SiteIF both proposes to the user a list of recommended documents (grouped according with the ADNKRONOS categories) and monitors the browsing of the user inside the site. For new users, it is possible to create an account. In this case SiteIF delays proposals recommendation after a couple of browsing sessions until the user model is sufficiently sizable. Figure 7 show a snapshots of the system while proposing some documents.

5. Evaluation

As far as evaluation is concerned, we focussed on two aspects: how much the reader's experience benefits by the adaptive access to the news web site, and how much the synset-based version of the system improves the previous (word-based) version (Stefani and Strapparava, 1998).

An adaptive way to access news should improve and make more attractive the reader's experience. We have compared personalized and non-personalized news

[4] See `http://www.ai.mit.edu/projects/iiip/doc/cl-http/home-page.html`

Figure 7. A snapshot of the SiteIF system showing some proposals to the user.

access under different situations. We alternated personalized and non-personalized news access for the same user on different days, both providing personalize access to a group of users and nonpersonalized access to another group. Our results show an increase in news readership by over 32% and 45%, respectively for word-based and synset-based versions, when headlines are sent in a personalized order to the user. These results are in agreement with other evaluation studies about the readership of personalized news (see for example (Billsus et al., 2002)).

The second experiment is more focused. We wanted to estimate how much the new version of SiteIF (synset-based) actually improves the accuracy of documents proposals with respect to the previous version of the system (word-based). However, setting a comparative test among user models, going beyond a generic user satisfaction is not straightforward. To evaluate whether and how the exploitation of the synset representation improves the accuracy of the semantic network modelling and filtering, we arranged an experiment whose goal was to compare the output of the two systems against the judgements of a human advisor.

We proceeded in the following way. First, a test set of about one hundred English news from the ADNKRONOS corpus were selected homogeneously with respect to the overall distribution in the ADNKRONOS categories[5] (i.e., culture, motors, etc...). The test set has been made available as a Web site, and then 12 ITC-irst researchers were asked to browse the site, simulating a user visiting the news site.

[5]Note that there is no explicit relation between the ADNKRONOS categories and the domains in WORDNET DOMAINS (see Section 2).

Users were instructed to select news article, according to their personal interests, to completely read it, and then to select another news, again according to their interests. This process was repeated until ten news were picked out.

After this phase, a human advisor, who was acquainted with the test corpus, was asked to analyze the documents chosen by the users, and to propose new potential interesting documents from the corpus. The advisor was an expert in sense annotation of large corpora. He was requested to follow the same procedure for each document set: documents were first grouped according to their ADNKRONOS category, and a new document was searched in the test corpus within that category. If a relevant document was found, it was added to the advisor proposals, otherwise no document for that category is proposed. Eventually, an additional document, outside the categories browsed by the user could be added by the advisor. On average, the advisor proposed 3 documents for a user document set.

At this point we compared the advisor proposals with the results of the two systems. To simulate the advisor behavior (i.e., it is allowed that for a given category no proposal is selected), all the system documents whose relevance was less than a fixed difference (20%) from the best document, were eliminated. After this selection, the system had about 10 documents for each user document set. Among them, the system selected the best scored document for each category, and eventually it proposed, on average, 3–4 documents.

Standard figures for precision and recall have been calculated considering the matches among the advisor and the systems documents. Precision is the ratio of recommended documents that are relevant, while the recall is the ratio of relevant documents that are recommended. In terms of our experiment we have $precision = \frac{|H \cap S|}{|S|}$ and $recall = \frac{|H \cap S|}{|H|}$, where H is the set of the human advisor proposals and S is the set of the system proposals.

Table III shows the result of the evaluation. The first column takes into account the document news, the second only the ADNKRONOS categories. We can note that precision considerably increases (34%) with the synset-based user model. This confirms the working hypothesis that substituting words with senses both in the modelling and in the filtering phase produces a more accurate output. The main reason, as expected, is that a synset-based retrieval allows to prefer documents with high degree of semantic coherence, which is not guaranteed in case of a word-based retrieval.

As for recall, it also gains some points (15%), even if it remains quite low. However, this does not seem a serious drawback for a pure recommender system,

Table III. Comparison between word-based UM and synset-based UM

	News		Categories	
	Precision	Recall	Precision	Recall
Word-Based UM	0.51	0.21	0.89	0.40
Synset-Based UM	0.85	0.36	0.97	0.43

where there is no the need to answer an explicit query (as it happens, for instance, in information retrieval systems), but rather the need is for an high quality (i.e., the precision) of the proposals.

6. Conclusions

We have presented a new version of SiteIF, a recommender system for a Web site of multilingual news. Exploiting a word sense based document representation, we have described a model of the user's interests based on word senses rather that on simply words. The main advantages of this approach are that semantic accuracy increases and that the model is independent from the language of the news.

To give a quantitative estimation of the improvements induced by a word sense based approach, a comparative experiment–sense-based vs. word-based user model–has been carried out, which has showed a significant higher precision in the system recommendations.

There are several areas for future developments. One point is to improve the disambiguation algorithms which are at the basis of the document representation. A promising direction (proposed in (Magnini and Strapparava, 2000)) is to design specific algorithms which consider the synset intersection of parallel news.

A second working direction concerns the possibility to develop clustering algorithms over the senses of the semantic network. For example, once the user model network is built, it could be useful to have the capability to dynamically infer some homogeneous user interest areas, and so have have a method to estimate how much the user model talks 'about homogeneous topics'. This would allow to arrange in uniform dynamic groups the recommended documents.

References

Artale, A., Magnini, B. and Strapparava, C.: 1997, WORDNET for Italian and its Use for Lexical Discrimination. In: *AI*IA97: Advances in Artificial Intelligence*. Springer-Verlag, pp. 346–356.

Ballesteros, L. and Croft, W. B.: 1997, Phrasal translation and query expansion techniques for cross-language information retrieval. In: N. J. Belkin, A. D. Narasimhalu, and P. Willett (eds.): *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-97)*, Vol. 31, special issue of *SIGIR Forum*. New York, pp. 84–91.

Billsus, D., Brunk, C. A., Evans, C., Gladish, B. and Pazzani M.: 2002, Adaptive interfaces for ubiquitous web. *Communications of The ACM* **45**(5), 34–38.

Brusilovsky, P.: 1998, Methods and Techniques of Adaptive Hypermedia. In: P. Brusilovsky, A. Kobsa and J. Vassileva (eds.): *Adaptive Hypertext and Hypermedia*. Dordrecht: Kluwer Academic Publisher, pp. 1–43.

Fellbaum, C.: 1998, *WordNet. An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fellbaum, C. and Vossen P.: (eds.): 2002, *Proceedings of the First International WordNet Conference.* Mysore, India.

Gonzalo, J., Verdejio, F., Chugur and Cigarran, J.: 1998a, Indexing with WordNet synsets can improve text retrieval. In: S. Harabagiu (ed.): *Proceedings of the Workshop 'Usage of WordNet in Natural Language Processing Systems'*. Montreal, Quebec, Canada.

Gonzalo, J., Verdejio, F., Peters, C. and Calzolari, N.: 1998b, Applying eurowordnet to cross-language text retrieval. *Computers and Humanities* **32**(2–3), 185–207.

Grefenstette, G.: 1998, *Cross-Language Information Retrieval,* Boston: Kluwer.

Hull, D. A.: 1997, Using Structured Queries for Disambiguation in Cross-Language Information Retrieval. In: *Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*. Stanford, CA, 73–81.

Kilgarriff, A. and Yallop, C.: 2000, What's in a thesaurus?. In: *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*. Athens, Greece, 1371–1379.

Krovetz, R. and Croft, W. B.: 1992, Lexical Ambiguity and Information Retrieval. *ACM Transactions on Information Systems* **10**(2), 115–141.

Lieberman, H., Dyke, N. W. V. and Vivacqua, A. S.: 1999, Let's Browse: A Collaborative Web Browsing Agent. In: *Proceedings of the 1999 International Conference on Intelligent User Interfaces*. pp. 65–68.

Magnini, B. and Cavaglià, G.: 2000, Integrating Subject Field Codes into WordNet. In: *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*. Athens, Greece, pp. 1413–1418.

Magnini, B. and Strapparava, C.: 2000, Experiments in Word Domain Disambiguation for Parallel Texts. In: *Proc. of SIGLEX Workshop on Word Senses and Multi-linguality*. Hong-Kong, pp. 27–33. held in conjunction with ACL2000.

Magnini, B. and Strapparava, C.: 2001, Improving User Modelling with Content-Based Techniques. In: *UM2001 User Modeling: Proc. of 8th International Conference on User Modeling (UM2001)*. Sonthofen (Germany), pp. 74–83.

Magnini, B., Strapparava, C., Pezzulo, G. and Gliozzo, A.: 2002, The role of domain information in word sense disambiguation. *Journal of Natural Language Engineering* **8**(4), 359–373.

Micarelli, A. and Sciarrone, F.: 2004, Anatomy and Empirical Evaluation of an Adaptive Web-Based Information Filtering System. *User Modeling and User-Adapted Interaction* (this issue).

Miller, G.: 1995, A lexical database for English. *Communications of the ACM* **38**(11), 39–41.

Minio, M. and Tasso, C.: 1996, User Modeling for Information Filtering on internet Services: Exploiting an Extended Version of the UMT Shell. In: *Proc. of Workshop on User Modeling for Information Filtering on the World Wide Web*. Kailia-Kuna Hawaii. Held in conjunction with UM'96.

Resnik, P.: 1995, Disambiguating Noun Groupings with Respect to WordNet Senses. In: *Proc. of third workshop on very large corpora*. MIT, Boston.

Schmid, H.: 1994, Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Stefani, A. and Strapparava, C.: 1998, Personalizing Access to Web Sites: The SiteIF Project. In: *Proc. of second Workshop on Adaptive Hypertext and Hypermedia*. Pittsburgh. Held in conjunction with HYPERTEXT 98. http://wwwis.win.tue.nl/ah98/Stefani/Stefani.html.

Stevenson, M. and Wilks, Y.: 2001, The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics* **27**(3), 321–350.

Strapparava, C., Magnini, B. and Stefani, A.: 2000, Sense-Based User Modelling for Web Sites. In: *Adaptive Hypermedia and Adaptive Web-Based Systems–Lecture Notes in Computer Science 1892*. Heidelberg: Springer-Verlag, pp. 388–391.

SENSEVAL-2: 2001. http://www.sle.sharp.co.uk/senseval2/.

Voorhees, E. M.: 1993, Using WordNet to disambiguate word senses for text retrieval. In: *Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Pittsburgh, Pennsylvania, pp. 171–180.

Vossen, P.: 1998, Special Issue on EuroWordNet. *Computers and Humanities* **32**.

Waern, A.: 2004, User involvement in automatic filtering an experimental study user involvement in automatic filtering–an experimental study. *User Modeling and User-Adapted Interaction*. (this issue).

Authors' vitae

Carlo Strapparava graduated in Computer Science at the University of Pisa with thesis in theoretical computer science in which some extensions of λ-calculi with strong type structures were studied.

He is currently a Senior Researcher at IRST in the Communication and Cognitive Technologies Division. His research activity covers artificial intelligence, natural language processing, intelligent interfaces, human-computer interaction, cognitive science, knowledge-based systems, user models, adaptive hypermedia, lexical knowledge bases (WordNet/MultiWordNet), word-sense disambiguation, and computational humour. He is author of over fifty published papers, both in scientific journals and in conference proceedings.

Bernardo Magnini is Senior Researcher at ITC-Irst (Istituto per la Ricerca Scientifica e Tecnologica) in Trento, Italy. He graduated at the University of Bologna with a thesis on Philosophy of Language. At ITC-Irst, he is involved in the Cognitive and Communication Technology division, where he coordinates a project on text processing technologies. His research interests are in the area of natural language processing and advanced information retrieval, with particular attention to question answering systems, word sense disambiguation and the application of NLP techniques to the Web scenario.