

Recommending tags for pictures based on text, visual content and user context

Stefanie Lindstaedt*, Viktoria Pammer*, †, Roland Mörzinger‡, Roman Kern*, Helmut Mülner‡, Claudia Wagner‡

*Know-Center

Inffeldgasse 21a, 8010 Graz, Austria.

Email: {slind,vpammer,rkern}@know-center.at}

†Knowledge Management Institute, Graz University of Technology

‡Joanneum Research

Steyrgasse 17-19 / Elisabethstrasse 20, 8010 Graz, Austria.

Email: {roland.moerzinger, helmut.muelner, claudia.wagner}@joanneum.at

Abstract—Imagine you are member of an online social system and want to upload a picture into the community pool. In current social software systems, you can probably tag your photo, share it or send it to a photo printing service and multiple other stuff. The system creates around you a space full of pictures, other interesting content (descriptions, comments) and full of users as well. The one thing current systems do not do, is understand what your pictures are about.

We present here a collection of functionalities that make a step in that direction when put together to be consumed by a tag recommendation system for pictures. We use the data richness inherent in social online environments for recommending tags by analysing different aspects of the same data (text, visual content and user context). We also give an assessment of the quality of thus recommended tags.

I. INTRODUCTION

Automatic tagging of multimedia content has long been an open issue in research. In the course of the semantic web, this issue gains new relevancy as it relies on the existence of formal content description. At the same time, there are more and more social software environments available online that contain huge amounts of multimedia, textual and user data. Currently, many users are enthusiastic enough to manually tag resources, which is a typical Web 2.0 approach that makes use of masses of users on the web. The goal of the scientific community must now be to bring science into the game. Clearly, such large systems would benefit from being able to automatically “understand” their own content. But even better, these systems possess an ideal data landscape for this challenge because they contain a wide variety of data that are interconnected.

The declared goal of this project was to create (semantic) structures based on the resources available in an online social software environment with multimedia content and to use these structures to automatically tag new multimedia content. We aimed at alleviating the problem of automatic tagging by including surrounding textual content and user information into the analysis.

In the work presented here we explore (the usefulness of) multiple views on one data set and we present a research

prototype tag recommendation system for pictures, the *tagr*. Our system performs a sort of data mash-up, based on scientifically challenging work going on in multiple areas like pattern recognition/image analysis, semantic web research and social network analysis / user modelling.

II. USER APPLICATION *tagr*

Imagine you are member of an online social system. You want to upload a picture into the community pool. The system analyses your picture and tries to guess the subject / topic of your picture. The system also presents to you similar pictures and similar users - you check them and see what tags others used. If they seem appropriate, you add them to your picture. While you add tags to your picture, the system reacts and suggests similar tags. - This is the idea behind *tagr*. In a real application, the picture and all collected tags would then need to be uploaded to the online social system.

Recommending tags for a given picture can be approached from two sides: looking for tags by analysing the picture and possibly already existing tags for it, or looking for tags by taking user preferences into account. The latter is the typical approach in collaborative filtering or recommendation systems (see [1]). The system we present here uses both paradigms. In the current implementation of the *tagr* however there is no exchange of information between the different functionality services. This will be one of the the main challenges in our continued work.

A. User interface

The *tagr*'s user interface is shown in figure 1(a). In the left corner you see the uploaded picture, and to its right side the tag-area with all tags that the user added to the picture. Below are three rows. The first is the picture row, showing similar pictures to the reference (left-most) picture (initially the picture you upload). Second is the tag row, showing tags similar to all tags in the tag-area. In the tag row, the *tagr*-user can take over tags by clicking on them. Third is the user row showing users similar to the reference (left-most) user (initially

the uploading *tagr*-user). The *tagr*-user can click on a picture (resp. user) to get a detail view as shown in figure 1(b). The detail view shows tags associated to the picture (resp. user) and lets the *tagr*-user choose this picture (resp. user) as reference for the picture row (resp. user row). Choosing a new reference changes the corresponding row. In the detail view, the *tagr*-user can also take over tags from the detailed picture (resp. user). The tag row changes when a new tag is taken over from the tag row or the detail view, or when the *tagr*-user manually enters a new tag into the tag-area.



(a) Main view.



(b) Detail view on selected picture (or user).

Fig. 1. Screenshot of tag recommendation system *tagr*.

B. System description

In general the *tagr* system is designed in a modular way, where it is potentially easy to reuse underlying services for different user applications.

A data service collects relevant data from a data source and makes it available to other services. Various functionality services analyse and re-organise data, providing added value / knowledge. A user application (*tagr*) consumes these functionality services.

The *tagr* relies on two data sources, namely WordNet¹ and Flickr². The data service grabs data from the Flickr-group “fruit & veg”³. The functionality services are an image similarity service, an image classification service, a WordNet term association service (which “owns” the WordNet database, i.e. it is the only service accessing it), a tag association service and a user similarity service.

Currently most of the analysis is done offline, i.e. the functionality services rely on a snapshot covering about 14K pictures. The data were aggregated into new structures that are queried at runtime, e.g. the tag association service creates a tag association index. This is a limitation of the current system

¹<http://wordnet.princeton.edu/>

²<http://www.flickr.com>

³<http://www.flickr.com/groups/fruitandveg/>

and reflects its status as a research prototype.

The data were grabbed from Flickr using the Flickr API. Only information that was publicly available was used.

The interaction between these services is sketched in figure 2.

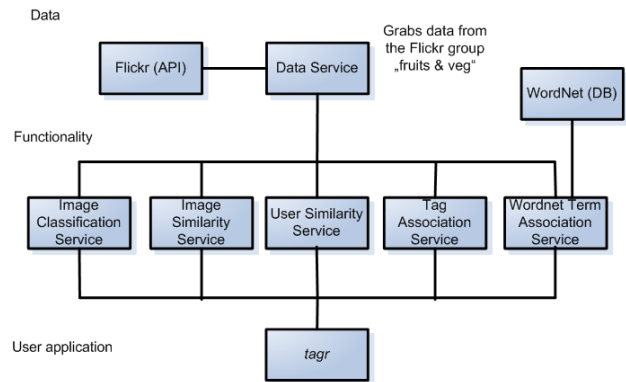


Fig. 2. Overview over data and functionality behind the *tagr*

III. SERVICES / BACKEND FUNCTIONALITY

A. Image classification

The image classification service automatically produces one or more tags for an un-tagged input image. For this purpose, a number of high-level tags of the “fruit & veg” folksonomy have been selected and classifiers for these tags have been learned. For the dataset used in the *tagr*, the most frequent tags which are not umbrella terms or subtopics of others were learned: ‘banana’, ‘blueberry’, ‘kiwi’, ‘orange’, ‘raspberry’, ‘strawberry’ and the negative class ‘other-fruits’. For each of these tags, a number of representative images (training set) was drawn from the “fruit & veg” image pool. Using image analysis, several image features were extracted, which basically describe global color and texture. For learning the classifier, a multi-class support vector machine was used. The classifier is able to classify “fruit & veg” images into the above mentioned classes. More details on the content-based features, the setup of this supervised learning technique and its evaluation (classification accuracy up to 71%) can be found at [2].

Automatically tagging an image is needed for kick-starting other services that require at least one tag/term as input, i.e. the tag/term association service. The image classification service feeds the tag-area to the right of the picture to-be-tagged (see figure 1(a)). Once one or more terms are automatically tagged, the user can use the connected services and browse through associated terms.

B. Image similarity

The goal of the image similarity service is to explore user-defined annotations of visually similar images (by cross-linking the image data with user annotations). Given an image, an arbitrary number of visually similar images can be

computed by content-based image retrieval. These images can be explored by the user, e.g. user-defined tags of the resulting images are propagated or similar images are requested again.

The content-based feature used for image similarity is the MPEG-7 Color Layout descriptor [2], which maps images into a high dimensional feature space. To achieve fast computation times when searching in this multidimensional space an efficient indexing techniques is required. For this purpose, we use a hybrid tree data structure and a non-linear distance measure in combination with Data Cartridge extension of the Oracle RDBMS. Experiments in large-scale databases of 20,000 images show that our similarity search service has a response time of 200 milliseconds on average.

The image similarity service feeds the picture row of the *tagr*, which shows the five most similar pictures to the current reference picture (initially the picture to-be-tagged).

C. Term associations based on Wordnet

WordNet is a lexical database for English (see [3]). It organises terms around *senses*, and a number of terms belong to each sense (*synset*), and describes hierarchical and non-hierarchical relations between senses. Our WordNet Term Association Service makes use of a Java interface⁴ to access the WordNet database. It offers the following functionality, directly re-using information described by WordNet: First, it can take a source set of terms and group them into senses, i.e. all terms in the source set which belong to one WordNet-sense are in one target set. Multiple group-membership is possible. Second, it can find all senses which one term represents and for each sense retrieves all synonyms. Third, it can find hyper-/hyponyms for a sense.

The main source of data covers only one topic (fruit). Clearly, functionality services can only aggregate existing data. Having only one domain-specific source of data therefore implies little to no support to a user who wishes to tag a picture of a different topic (or a so-far unknown fruit). Integrating a second, more general, source of data alleviates this problem.

The *tagr* uses the second and third functionality of the WordNet term association service, namely finding all senses of one term, all synonyms that go with the senses as well as hypernym-senses and their synsets. Within the *tagr*, this service feeds the tag row together with the tag association service. For application in the tag row, each group of words gets ranked higher for every term in it that is also a result of the tag association service.

D. Tag association based on statistical distribution

The tag association service is based on the statistical distribution of tags in relation to the photos. Similar tags are found using a co-occurrence, or better a co-tagging, analysis. The critical aspect of this part is to select an appropriate algorithm that produces results (associated tags) that closely match the data set.

The data service is queried to retrieve all available photos together with all their tags. This information is transformed into an index that contains precomputed similarity weights between tags. This index represents a tag-tag matrix similar to the term-term correlation matrix presented in [4]. This matrix consists of input tags as rows and output tags as columns and each element holds the precomputed similarity of two tags based on their usage distribution. To find similar tags, a search within the similarity index is executed with one or more tags as input and a sorted, weighted list of tags as output. These tags are then used for recommendation.

The similarity of one tag (*i*) to another (*j*) is calculated using the number of shared photos ($sharedPhotos_{i,j}$) in relation to the number of times this tag is used using the formula $w_{i,j} = \frac{sharedPhotos_{i,j}}{mean(photoCount_i, photoCount_j)}$, whereas $photoCount_i$ is the number of photos tagged with *i*. This similarity measure was determined experimentally to be the best for this setting. Similar to Deshpande and Karypis in [5] we considered various alternatives, such as the cosine similarity $w_{i,j} = \frac{tag_i^T tag_j}{||tag_i|| ||tag_j||}$ or $w_{i,j} = \frac{sharedPhotos_{i,j}}{photoCount_i}$ as well as using *max* instead of the *mean* function, but they did not offer an improved performance.

For a single tag the result set of the recommended tags is the list of all tags that share at least one photo, sorted by their similarity weight. If more than one tag is input the result sets for each single tag are merged and the weights are calculated using $w_{out} = \sum_{in} w_{in,out}$. As all similarity weights are already computed at indexing time, this operation is relatively fast.

Within the *tagr* this service feeds the tag row together with the WordNet term association service. A tag gets additional points for ranking if it also occurs in the result set of the WordNet term association service.

E. User similarity

An important feature of human behaviour is the tendency to consume a limited set of items (in our case tags) within one group. So a set of items becomes characteristic for a group of similar users [6]. This observation motivates our approach to recommend tags to a user based on tags given by similar users in a sufficiently similar context.

Recommendation systems in general are described as systems whose task it is to “estimate ratings for items not yet seen by the user”, as Adomavicius and Tuzhilin express in [7]. Adopting their categorisation of recommendation systems, the user similarity service, as well as the tag association service actually, correspond to a collaborative recommendation system, recommending tags based on what tags similar users used in a similar context (same Flickr group).

We identified four groups of relations between users:

- Personal relations: A user knows another user personally from real life or online life. Example: contacts in contact list.

⁴<http://www.mit.edu/markaf/projects/wordnet/>

- Relations via objects of sociality (pictures): If user A knows user B via this relations, A knows at least one picture of B. Examples: comments on the same picture, voter/owner of favourite picture.
- Common ground relation: If two users share a common ground, they are similar to each other. Example: membership in same group, same or overlapping interests, same hometown.
- Common behaviour relation: This reflects the similarity of users with regard to the level / kind of activity in a social context. Example: similar number of comments and testimonials, similar responsibilities like admin/moderator.

The described features form a feature space which can be considered as a directed (not all relations are bidirectional) graph, where the nodes are users and the edges describe the relations between the nodes. Currently the direction of relations is discarded, but we intend to integrate this additional information in future developments using techniques similar to [8].

The more features two users have in common, the more similar they are. Clearly, the features mentioned above do not have the same importance to the social distance of users. We distributed weights heuristically, giving higher weights to personal relations and relations via objects of sociality. The enumeration of relations above actually reflects the chosen ranking. Similarity between users is then calculated by the weighted Manhattan distance of the feature vectors.

The user similarity service feeds the user row within the *tagr*. To each of the five similar users shown in the user row, his/her most frequent tags in the current social context (=Flickr group “fruit & veg”) are given.

IV. EVALUATION

The perception of usefulness for tag recommendation of the *tagr* has first been evaluated in a user study. The goal was to find out whether users feel supported by the offered functionality in tagging a picture. Additionally we studied whether and how manually given tags differ from *tagr*-supported given tags. Then, the objective appropriateness of each functionality service with regard to tag recommendation was evaluated in a separate study. Results of both evaluations are summarized here.

a) Study 1: For the user study, a set of 40 pictures were taken from the Flickr group “fruit & veg”. These pictures are distinct from the set of pictures on which our backend services rely. However, no new Flickr-users were introduced to our system. 4 test users tagged these pictures. Every test user had to manually tag 20 pictures and use the *tagr* to tag 20 more pictures. Each picture was tagged at least once manually and at least once with the support of *tagr*. Afterwards, each of the test users was asked for their opinion regarding usefulness, joy of use and potential other applications.

Let Z_i be the set of tags given manually by all test users and

V_i the set of tags given with support of the *tagr* by all test users for the i -th picture. Then tag-ratio $\sum_i |Z_i|/|V_i|$ is the proportion of the number of manual tags to supported tags. In the performed user study the tag-ratio is ≈ 1.1 .

b) Study 2: Next, 535 new pictures were taken from the Flickr group “fruit & veg”. In this study, each of the functionality service was used for predicting a number of tags. The reference tags in this study were the original Flickr tags.

In this study precision and recall of each of the functionality services was measured where the precision is $p = |M|/|V|$ and the recall is $r = |M|/|Z|$ with M the intersection between Z (set of tags given by Flickr user) and V (set of tags reachable by *tagr*). We call the following tags reachable by the *tagr*

- Picture row: 5 pictures are shown, tags attached to any of these pictures (corresponds to 1-click-away from the *tagr* user)
- Tag row: knowing one tag of the target tag set Z , the recommended tags. This has to be iterated (“leave-one-out”) for all tags in Z
- User row: 5 users are shown, tags related to any of these users (corresponds to 1-click-away from the *tagr*-user)

The precision and recall values for each functionality are:

Picture row	$p = 0.05$	$r = 0.14$
Tag row	$p = 0.08$	$r = 0.17$
User row	$p = 0.03$	$r = 0.18$

c) Synthesis of results: The high tag-ratio (above 1) means that on average a picture gets more different tags if it is manually tagged than if it is tagged with the *tagr*. Obviously, using an automated support leads to unification of vocabulary.

The precision values are obviously quite low. We think the reason for this is mainly that separate information (e.g. from image analysis and tag association) is not yet connected. The low recall is more difficult to explain. While analysing single pictures, we got the impression that most pictures have around five or six tags, which means that on average we can recommend two tags. The not-reachable tags usually were related to something not directly inferable from the picture (i.e. specific location or event).

The test users felt most supported by the tag row, which is reasonable as the tag row gives the most direct support for the task of finding appropriate tags. They felt conditionally supported by the picture row, depending on how similar the retrieved images actually were to the picture to-be-tagged. From a technical point of course, the system here is limited by what pictures are available in the data set, while the tag row was able to provide tags also from outside the domain-specific data set by using WordNet. The test users felt least supported by the user row. Our interpretation is that as our test users had to assume the identity of a Flickr user for testing they could not easily relate to the persons found similar. They agreed that

the user row provided a different point of view. Interestingly however, the test users were quite clear in ranking the usefulness of the functionalities (tag row, picture row and last user row), whereas the objective difference in both precision and recall are not marked.

Also, despite the low precision and recall values test users did find the *tagr* useful. This might be because they do not actually reach *all* tags, but will click only on the most relevant picture or user. Or they will enter a very relevant tag and look for associated tags.

It is worth noting that there is a difference between the results from these two studies and the results from another study using the same image similarity described in [2], where precision is given as higher. The reason for the difference generally lies in slightly different test set-ups w.r.t. to which tags are propagated. In the set-up in [2] more similar pictures (fifty instead of five) were used for tag propagation, but only those tags that occurred multiple times were counted as “recommended” (versus counting all tags for all five similar pictures as done here). Additionally, knowledge about semantic relationship (hyponymy) was used to select tags. Also the judgement of correct versus wrong tags was implemented differently: here correct means “tag given independently by another user” whereas in [2] test users judged automatically propagated tags.

Concluding and looking out to future work and ways for improvement, the differences in result indicate that taking into account a larger number of items of one dimension (e.g. pictures, users, tags) and increasing post-processing on the resulting set of tags leads to an improved precision. Additionally, we assume that combining knowledge from multiple dimensions (pictures, tags, users) will also improve quality.

V. RELATED WORK

We address related work regarding a number of different aspects of our work. We specifically discuss related work that at the same time shows the way for future improvements and work.

Our work has so far not been concerned with attaching formal semantics to tags. Rather we do the inverse and use available semantic models (WordNet) to propagate tags. However, going from informal (tags, folksonomies) to formal content description is the obvious continuance of our work.

Schmitz [9] creates a subsumption hierarchy of the Flickr tags by using the distribution and co-occurrence properties of tags. This could well serve as complementary approach to using WordNet for creating a hierarchy.

An interesting work by Rattenbury et al. specializes in detecting whether a (Flickr) tag corresponds to an event or to a place (see [10]). The underlying hypothesis is that event-related tags have significant peaks in the time-distribution whereas place-related tags have significant peaks in their location distribution. Their results support this hypothesis. Sense disambiguation is a long-standing problem in artificial

intelligence, specifically natural language processing. For a general survey on its history and different principles we refer the interested reader to Ide and Véronis[11]. Specifically interesting in relation to our work is of course the use of WordNet as reference knowledge source, as e.g. presented by Li et al. in [12]. The idea of Li et al. is to use the verb determining a noun object within a sentence to disambiguate the noun.

In a tag recommendation system, an underlying ontology could be used to display tags in different categories (higher-level concepts). This potentially enables users to better focus on tags that seem relevant to them. An available hierarchy would make automatic disambiguation desirable and thus enable real “recognition” of topic by the system, e.g. “this picture is most probably about the fruit Kiwi and not about the bird”.

The relevant context for disambiguation then consists potentially of already given tags, similar items (pictures in a tag recommendation system for pictures) and similar users. All this must be addressed by further work, if the goal is indeed to “understand” pictures and describe them semantically.

We round off this section by mentioning some Flickr mash-ups. These are just a snapshot of all available mash-ups, and we focus on those image search systems that provide services which resemble our image search interface.

Flickr Suggestions⁵ is an online Flickr mash-up that allows browsing through Flickr data. Flickr Suggestions has three basic functionalities: browsing along visually similar pictures, browsing along visually similar pictures with selected tags and browsing along similar users. Picture similarity is based on wavelet decompositions of images⁶, user similarity is apparently calculated based on common favourites and “submitted photos”. Unfortunately it is not clear, whether submitted photos are compared via their tags or via image-similarity metrics. *retrievr*⁷ is another Flickr mash-up that uses wavelet decomposition to find similar pictures. It offers the possibility to either draw a sketch or upload a picture and search for similar pictures.

TagTree⁸ is a tag-based exploration of Flickr images. Starting from one tag, co-occurring tags are displayed. By repeatedly clicking on tags, a tag hierarchy is created. At each level, clicking on a tag also shows pictures tagged with the selected and all higher-level tags. Similar to the TagTree is the Flickr related tag browser⁹. Given a tag it shows picture results and similar tags. It is unclear how similar tags are calculated.

Most Flickr mash-ups create a very good user experience, providing playful and fun user interfaces. Unfortunately not all of them detail the underlying algorithms. The challenge for the (semantic) web scientific community now lies in adding

⁵<http://fs.imgseek.net/>

⁶<http://server.imgseek.net/category/documentation/architecture/>

⁷<http://labs.systemone.at/retrievr>

⁸<http://www.tagtree.net>

⁹http://www.airtightinteractive.com/projects/related_tag_browser/app/

theoretically well-grounded algorithms formal semantics to Web2.0 data and discussing algorithms for effectively mashing up and interpreting data. At the same time, the standard for user experience is set high by Web2.0 applications, and our estimation is that also scientific work will be measured against this standard.

VI. CONCLUSION

We have developed functionality for a tag recommendation system for pictures that depends on rich data, interlinking pictures with textual descriptions and user data. We have also developed a specific research prototype, the *tagr*, which makes use of data from the Flickr group “fruit & veg” and of the electronic lexical database WordNet.

In order to effectively profit from the variety of dimensions available in the underlying data (pictures, words, users), techniques from multiple fields were used to develop the collection of functionality services, including image analysis, statistical text analysis and social network analysis. We are strongly of the impression that this is one (albeit small) step ahead of state of the art, in that it combines many dimensions and uses scientifically well-grounded methods for performed calculations (which is often not the case for similar mash-ups). Still, we have also discussed (in section IV), that the quality of tag recommendations of our system is still far from being perfect. We assume that one reason is that we do not yet really *mix* the dimensions of text, visual content and user preferences for the purpose of tag recommendation. This is clearly the next logical step for the near future. A little bit farther ahead seems the goal of semantic recognition of a picture, which is strongly related to the very up-to-date research on the joining of folksonomies and ontologies.

ACKNOWLEDGMENT

The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria.

REFERENCES

- [1] J. S. Breese, D. Heckerman, and C. Kadie, “Empirical analysis of predictive algorithms for collaborative filtering,” in *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, 1998, pp. 43–52.
- [2] R. Mörzinger, R. Sorschag, G. Thallinger, and S. Lindstaedt, “Automatic image annotation using visual content and folksonomies,” in *Proceedings of the Metadata Mining for Image Understanding Workshop at VISAPP2008*, Funchal, Madeira, Portugal, January 2008.
- [3] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*. The MIT Press, May 1998.
- [4] R. A. Baeza-Yates and B. A. Ribeiro-Neto, *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.
- [5] M. Deshpande and G. Karypis, “Item-based top-n recommendation algorithms,” *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 143–177, 2004.
- [6] M. F. Trujillo, M. Millán, and E. Ortiz, “A recommender system based on multi-features,” in *ICCSA (2)*, ser. Lecture Notes in Computer Science, O. Gervasi and M. L. Gavrilova, Eds., vol. 4706. Springer, 2007, pp. 370–382.

- [7] G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.
- [8] J. Duch and A. Arenas, “Community detection in complex networks using extremal optimization,” *Physical Review E*, vol. 72, p. 027104, 2005.
- [9] P. Schmitz, “Inducing ontology from flickr tags,” in *Proceedings of the Collaborative Web Tagging Workshop at WWW2006*, Edinburgh, Scotland, May 2006.
- [10] T. Rattenbury, N. Good, and M. Naaman, “Towards automatic extraction of event and place semantics from flickr tags,” in *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference*. New York, NY, USA: ACM Press, 2007, pp. 103–110.
- [11] N. Ide and J. Véronis, “Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art,” *Computational Linguistics*, vol. 24, no. 1, pp. 1–40, 1998.
- [12] X. Li, S. Szpakowicz, and S. Matwin, “A WordNet-based algorithm for word sense disambiguation,” in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95*, Montréal, Québec, Canada, August 20-25 1995, pp. 1368–1374.