Usage based indexing of web resources with natural language processing

Armelle Brun, Anne Boyer

USAGE BASED INDEXING OF WEB RESOURCES WITH NATURAL LANGUAGE PROCESSING

Armelle Brun, Anne Boyer
INRIA Lorraine - Université Nancy2 , France
anne.boyer@loria.fr, armelle.brun@loria.fr

Keywords: Recommender systems, collaborative filtering, analysis of usage, statistical language model, resource indexing

Abstract: Due to the huge amount of available information via Internet, the identification of reliable and interesting items becomes more and more difficult and time consuming. This paper is a position paper describing our intended work in the framework of multimedia information retrieval by browsing techniques within web navigation. It relies on a usage-based indexing of resources: we ignore the nature, the content and the structure of resources. We describe a new approach taking advantage of the similarity between statistical modeling of language and document retrieval systems. A syntax of usage is computed that designs a Statistical Grammar of Usage (SGU). A SGU enables resources classification to perform a personalized navigation assistant tool. It relies both on collaborative filtering to compute virtual communities of users and a new distance dependent trigger model. The resulting SGU is a community dependent SGU.

1 INTRODUCTION

On Internet the identification of reliable and interesting items becomes more and more difficult and time consuming, even for skilled people using dedicated tools, such as powerful search engines. Due to the huge amount of online resources, the major difficulty is nevermore to know if a pertinent document is available but to identify the more reliable and interesting items among the overwhelming stream of available information. A key factor of success in information retrieval and delivery is the development of powerful tools easy-to-use for a large audience.

Different approaches for resources retrieval are explored, such as content analysis, keywords indexing and identification, topic detection, etc. (Baeza-Yates and Ribeiro-Neto, 1999). A major difficulty inherent to such approaches is that one keyword may have different meanings, or not, dependent of the user, his/her context and the history of his/her past navigations. Moreover two different keywords may have similar meanings, depending on the context. Expressing a query is a difficult task for many people and a lot of research and industrial projects deal with query assistance. Furthermore automatic indexing of multimedia resources is still a hard research problem. To cope with these difficulties we decide to investigate another way by ignoring the content, the nature, the format and the structure of resources.

This position paper describes our intended work, relying on past researches both on collaborative filtering (Castagnos and Boyer, 2006a; Castagnos and Boyer, 2006b) and statistical language modeling (Smaïli et al., 1999; Brun et al., 2002). We aim at providing a new web browsing tool based on an analysis of usage. This tool enables multimedia information retrieval by browsing techniques without expressing any query: users are modelized without requiring any preferences elicitation. This approach manages easily heterogeneous items (video, audio, textual, multimedia) with a single treatment, as classical methods require dedicated tools for resource tagging.

We plan to extract frequent patterns of consultations by taking advantage of the analogy between language-based statistical modeling and resource retrieval. These frequent patterns allow the design of syntax of usage, relying on the hypothesis that there is logic and coherency defining implicit ”rules” inside a navigation. The resulting Statistical Grammar of Usage enables a classification, clustering and selection

of resources to design personalized filtering.

In the next section, the problem of retrieving resources when browsing is stated and our approach based on the use of statistical language models is detailed. The following section presents the most popular statistical language models and their appropriateness to web browsing. An adaptation of the trigger language model is introduced. Section 4 puts forward the community-based Statistical Grammar of Usage. Discussion and perspectives conclude this paper.

2 Our approach

Our web browsing tool helps users during a navigation process: it suggests the pertinent items to a specific user, given his/her past navigation and context. The aim is to compute the **pertinence** of any resource. The pertinence of a resource is the interest of a user for it and allows to compute **predictions** of resources (the highest the pertinence of a resource is, the highest is its probability to be suggested).

First, we hypothesize an **implicit search**, it means that the active user has no explicit queries to formulate. Secondly, we consider as a **consultation** the sequence of one or more items, dedicated to a given search. A multi-navigation is the mix of different consultations within a single browsing process. One single consultation is called a mono-navigation.

A **resource** is any item (textual, audio, video or multimedia document, web page, hyperlink, forum, blog, website, etc.), viewed as an elementary and indivisible entity without any information about its format, its content or any semantic or topic indexing. The only data describing *a priori* a resource is a normalized mark called **identifier**, enabling to identify and to locate it.

Our approach relies on an analysis of usage. A **usage** is any data, explicitly or implicitly left by the user during navigation. For example, history of consultation, click-stream or log files are implicit data about the interest of the visited items for the active user. We call **appreciation** any measure of the user's satisfaction. This measure can be either an explicit information as votes, annotations or any estimation computed from implicit data (Chan, 1999).

An advantage of our approach is that it only takes into account a measure of the user's interest for a given resource, which is directly linked to the pertinence criterion: the user's satisfaction. Our approach computes a personalized indexing of resources not in terms of its intrinsic nature but in terms of a more subjective but more reliable and pertinent criterion, i.e. the user's context, preferences and habits. Then this approach manages heterogeneous resources with a single treatment.

The question is how to estimate the *a priori* pertinence of a resource for a given user. The difficulty relies on sparsity of data: we don't have any appreciation of a resource if this user has not seen it and usually most of resources have not been seen by this user. As we design a personalized tool, the pertinence cannot be a context independent measure (the **context** is both the user's profile and history).

To compute the *a priori* pertinence of a resource, we plan to design a grammar of usage. As a **grammar of language** is the set of rules describing the relation between words, a **grammar of usage** is the set of rules describing the relation between resources. A grammar of language estimates if a word is pertinent given the beginning of a sentence. A grammar of usage allows to estimate if a resource is relevant for a specific user given his/her previous consultations.

There is no *a priori* grammar of usage, as Internet is a dynamic and moving environment. A means to cope with the difficulty of designing an *a priori* grammar is the use of a statistical approach based on usage analysis. As huge usage corpora are available (log files and clickstream) it makes it possible to explicit regularities in terms of resource consultations. This statistical approach can be investigated in a similar way to language modeling based on statistical models (defining a **Statistical Grammar of Language**).

The resulting grammar is called a **Statistical Grammar of Usage** (SGU). It enables the computation of the probability of a resource given the active user and his/her sequence of navigation. This probability measures the pertinence of the resource.

A SGU, if trained on the whole usage corpus, is a general grammar since it is learned for all users in all contexts. The accuracy of such a grammar is insufficient and furthermore, the presupposed logic and coherency between users becomes a too strong and unrealistic hypothesis. Given two users, it seems unlikely that they exhibit the same resource consultation behavior: the SGU has to be personalized. Nevertheless, learning a user-specific SGU requires a large amount of data for each user and it is unrealistic to wait for collecting enough data to train it. It is the reason why we determine groups of users with similar behavior called **communities**. We compute a SGU for each community and design a **community-based SGU**. This approach is used in statistical language modeling where corpus is split into topic sub-corpora.

Users are preclassified into a set of coherent communities, in terms of resource consultation behavior. Collaborative filtering techniques are a challengeable means to build coherent communities in terms of

usage. The principle of collaborative filtering techniques (Herlocker et al., 2004) amounts to identifying the active user to a set of users having the same tastes and, that, based on his/her preferences and past visited resources. This approach relies on the hypothesis that users who like the same documents have the same topics of interests. Another is that people have relatively constant likings. Thus, it is possible to predict resources likely to match user's expectations by taking advantage of experience of his/her community.

A first comment on usual collaborative filtering techniques is that the structure of navigation is ignored. However, this aspect can be crucial in some applications such as web browsing. For example, a user may not like a resource because he/she has not previously read a prerequisite resource. Thus the SGU submits a resource when it becomes pertinent for a user, for example when he/she has read all prerequisites. As statistical language models emphasize the order of words in sentences, it seems interesting to determine if such models and collaborative filtering can be used together to improve the quality of suggestions.

3 Statistical language models

3.1 Overview

The role of a statistical language model (**SLM**) is to assign a likelihood to a given sentence (or sequence of words) in a language (Jelinek and Mercer, 1980; Rosenfeld, 2000). A SLM is defined as a set of probabilities associated to sequences of words. These probabilities reflect the likelihood of those sequences.

SLM are widely used in various natural language applications such as optical character recognition, automatic speech recognition, etc.

Let the word sequence $W = w_1, \ldots, w_S$. The probability of W is computed as the product of the conditional probabilities of each word w_i in the sequence:

$$P(W) = \prod_{i=1}^{S} P(w_i \mid h_i) = \prod_{i=1}^{S} P(w_i, \mid w_1 \ldots, w_{i-1}) \quad (1)$$

where w_i is the i^{th} word of W. h_i is the history of w_i. To estimate these probabilities, a vocabulary $V = \{w_j\}$ is stated. The probability of sequences of words are trained on a text corpus, the training corpus.

3.2 Advantage of SLM for web browsing

Web browsing and statistical language modeling domains seems similar in several points. First, statistical language modeling uses a vocabulary made up of words. This set can be viewed as similar to the set of resources R of the web. Then, the text corpus is made up of sentences of words, they can be viewed as similar to the sequences of consultations of the usage corpus. A sequence of S words in a sentence is similar to a sequence of consultation of S resources. Finally, the presence of a word in a sentence mainly depends on its previous words, as the consultation of a resource mainly depends of the preceding consultations.
Given these similarities, we can naturally investigate the exploitation of models used in statistical language modeling into a web browsing assistant. As noticed in the previous section, these models have the characteristic that the order of the elements in the history is crucial. This aspect may be important for specific resources in web browsing.

However, we have to notice that web browsing and natural language processing have two major differences. The first one is that it is possible that a user may mix different queries within a single history ("multi-navigation") but it is unrealistic to mix different sentences when speaking or writing. This first remark brings us to consider a generalization of SLM to take into account "multi-navigation" in the browsing process. The second one is that natural language exhibits strongest constraints: each word in a sentence is important and deleting or adding a word may change the meaning of the sentence. Web browsing is not so sensitive and adding or deleting a specific resource within a navigation may have no impact. Then we have to consider permissive models, able to manage less constrained histories.

3.3 n-grams language models

Due to computational constraints and probability reliance, the whole history h_i of a word w_i cannot be systematically used to compute the probability of W. Classical SLM aim at reducing the size of the history while not decreasing performance.

n-grams models (Jurawski and Martin, 2000) reduce the history of words to their $n-1$ previous words. These models are the most commonly used in most of natural language applications. The probability of a given word w_i given history h_i is computed as follows:

$$P(w_i|w_{i-n+1} \ldots w_{i-1}) = \frac{N(w_{i-n+1} \ldots w_{i-1}, w_i)}{N(w_{i-n+1} \ldots w_{i-1})} \quad (2)$$

where $N(.)$ is the number of occurrences of the argument, in the training corpus.

n-grams model can be directly used in web browsing assistance. In the previous section, we put forward that the quality of the model will be increased if it is dedicated to a community and trained on the corresponding community usage corpus. Thus, the usage corpus is split into community usage corpora and a model is trained on each community corpus.

Let a community c_j and a sequence of consultations of resources $h_j = R_{j1},\ldots,R_{ji-1}$. The n-grams model computes the probability for each resource $R_i \in R$:

$$P_n(R_i \mid R_{i-n+1},\ldots,R_{i-1},c_j) = \frac{N_{c_j}(R_{i-n+1},\ldots,R_{i-1},R_i)}{N_{c_j}(R_{i-n+1},\ldots,R_{i-1})} \quad (3)$$

where $N_{c_j}(.)$ is the number of occurrences of the parameter in the community usage corpus c_j. The history h_j has been reduced to the $n-1$ last resources consulted, other resources are discarded. Thus, this model assumes that the consultation of a resource R_i does not mainly depend on resources consulted far from R_i.

As previously mentioned, adding or deleting a resource in a sequence of consultations has a lower influence on the result of the search than adding or deleting a word in a sentence. Thus, this model does not ideally match our retrieval problem since the history considered is the exact sequence of consultations $R_{i-n+1}\ldots R_{i-1}$, that may be too restrictive in the general case. However, this model may be suitable for frequent sequences of consultations, that can be considered as "patterns of consultation". They are assigned a high probability, thus increasing the probability of resources inside such sequences. It should be interesting to take into account, in a more adequate way, such "patterns of consultations".

As n-grams models exhibit strong constraints, we are also interested in more permissive models. Trigger-based language models seem to me more adequate to less constraint histories such as navigation.

3.4 Trigger-based language models

Trigger-based models (Rosenfeld, 1996) aim at considering long-time dependence between two words (w_x and w_y for instance). Dependence is measured by Mutual Information (MI) (Abramson, 1963). This measure can easily integrate long-time dependence by using a distance parameter d. d is the maximum number of words occurring between w_x and w_y, a window of d words is thus considered. MI between words w_x and w_y, in a window of d words, is computed as:

$$MI(w_x,w_y,d) = log\frac{P_d(w_x,w_y)}{P_d(w_x)P_d(w_y)} \quad (4)$$

where $P_d(w_x,w_y)$ is the probability of w_x preceding w_y at a distance at most d, in the training corpus.

A couple (w_x,w_y) with a high MI value means that w_x and w_y are highly correlated and the presence of w_x raises the probability of occurrence of w_y, at a maximal distance of d words. (w_x,w_y) is named a trigger. This model considers only highly correlated pairs of words (corresponding to high MI values), useless pairs are discarded.

In our web browsing assistant tool, the trigger model is made up of triggers of resources (R_x,R_y). The consultation of R_x triggers the consultation of R_y, at a maximal distance of d resources. As MI measure is not symmetric ($MI(R_x,R_y) \neq MI(R_y,R_x)$), this model integrates order between resources, that may be crucial for specifi c resources.

The advantage of this model is the long-time dependence between resources. In a consultation, two resources can be viewed with various values of distance without changing the meaning of the consultation. Trigger models enable to modelize this kind of influence, when the distance between items is not discriminant but the order of occurrence is meaningful. Such a model is less constrained than n-grams models and seems to be adequate to the navigation problem.

Similarly to n-grams model, a trigger-model is developed for each community c_j. MI values are computed for each couple of resources and for each community. A set of most related triggers is extracted for each community c_j. This set is called S_{c_j}.

The probability of a resource R_i, given the community c_j, its corresponding set of triggers S_{c_j} and the sequence of consultation of resources $h_j = R_1,\ldots,R_{i-1}$ is:

$$P_t(R_i \mid h_j,c_j) = \frac{\sum_{R_x \in h_j} \delta_{R_x,R_i,h_j,S_{c_j}}}{\sum_{R_x \in h_j}\sum_{R_y \in R} \delta_{R_x,R_y,h_j,S_{c_j}}} \quad (5)$$

with

$$\delta_{R_x,R_i,h_j,S_{c_j}} = \begin{cases} 1 & (R_x,R_i) \in S_{c_j} \text{ and } d_j(R_x,R_i) \leq d \\ 0 & otherwise \end{cases}$$

$d_j(R_x,R_i)$ is the distance between R_x and R_i in h_j.

3.5 Distance-dependent trigger model

State of the art trigger models, as previously presented, aim at considering long distance relations (distance between 0 and d). Each couple of resources appearing "frequently" at a maximal distance d is set in the model.

However, this kind of relation between resources is too general and we assume that fi ner relations are present in the corpus. Let two resources R_x and R_y, always cooccurring at a maximal distance u where

$u << d$. During training, each cooccurrence (at a distance at most u) of this couple is taken into account as a cooccurrence of this couple at a maximal distance d, the corresponding MI value is computed. If this MI value is high, the pair is selected as a trigger pair. During test, if R_x occurs, the probability of R_y is increased while d resources have not been consulted. This trigger is misused: R_x and R_y appear at a distance at most u during training, this distance u should also be considered during test: the probability of R_y should be increased at a distance lower than u.

Thus, we assume that taking into account fi ner relations by using the actual training distance between R_x and R_y will correspond to a better modelization of relations between resources and thus increase the quality of the trigger model.

We propose a model able to consider several kinds of relations, such as

1. The above relation: two resources are mainly consulted at a distance lower than u with $u \leq d$.
2. The converse relation : two resources are mainly consulted at a distance larger than l where $0 \leq l$
3. Two resources are mainly consulted at a distance between l and u with $l \leq u \leq d$

However fi xing, for all triggers, the same value for l and the same value for u can be suboptimal: obviously these values depend on both resources of the trigger.

3.5.1 Computing optimal values for l and u

Given a community c_j and a pair (R_x,R_y), the optimal values of l^* and u^* are the ones maximizing:

$$l^*, u^* = \underset{l,u}{argmax} \quad MI_{c_j}(R_x, R_y, l, u) \tag{6}$$

where l and u rank from 0 to d. $MI_{c_j}(R_x, R_y, l, u)$ is the mutual information of resources R_x and R_y at a distance ranking from l to u in the community c_j and is computed as follows:

$$MI_{c_j}(R_x, R_y, l, u) = \log \frac{P_{c_j,l,u}(R_x, R_y)}{P_{c_j,l,u}(R_x)\, P_{c_j,l,u}(R_y)} \tag{7}$$

$P_{c_j,l,u}(R_x, R_y)$ is the probability of cooccurrence of R_x and R_y at a distance ranking from l to u in the community corpus c_j.

Let us notice that the MI value is not reliable if values in the denominator are low. Indeed, when those values are too low, the MI value is anormally highed, then does not represent the real correlation value between the two resources. Thus, MI will not be computed for pairs with low denominator values.

3.5.2 Formalization of the new trigger model.

The trigger model we propose here is made up of highly correlated and distance-dependent pairs: a lower value of distance l and an upper value of distance u are considered for each pair. Given the trigger (R_x, R_y, l, u), the probability of R_y is increased if R_x occurs in the history h_j, and the distance $d_j(R_x, R_y)$ between R_x and R_y is between l and u.

Thus, given the community c_j, the corresponding set of triggers S_{c_j} and a history h_j, the probability assigned to a given resource R_i by the trigger model is defi ned as in equation (5) and

$$\delta_{R_x,R_i,h_j,S_{c_j}} = \begin{cases} 1 & (R_x, R_i) \in S_{c_j} \ \textit{and} \\ & l \leq d_j(R_x, R_i) \leq u \\ 0 & \textit{otherwise} \end{cases}$$

Both n-grams and distance-dependent triggers models are candidates to integration into a web browsing tool. A n-grams model computes the probability of sequences of consultation, a trigger model extracts pairs of distant resources. Consequently, both are interesting to achieve our goal and will be integrated in the community-based SGU we propose.

4 Towards a community-based SGU

The SGU we propose has the advantage of considering both the community of the active user and his/her consultation history, whereas state of the art models usually exploit the set of consultations. The use of this model relies on two steps:

1. Determination of the community c_j of user U_j.
2. Computation of the probability of each resource R_i, given c_j and the history h_j of U_j.

4.1 Community determination

The objective is to compute a set of user communities based on an analysis of usage. To achieve this goal, we use collaborative fi ltering techniques. The set of users is split into classes by using a recursive k-means like algorithm (Castagnos and Boyer, 2006a), the similarity between two users is estimated as the mean of the distance for each commonly voted resource.

The whole corpus is then split into a set of community sub-corpora. Each community corpus is made up of usage of any user in the community. A user is then assigned to the closest community using the same similarity measure.

4.2 Probability computation

Given the community c_j of user U_j, and his history h_j, the computation of the probability of a resource R_i relies on three sub-models based on language models presented in section 4.

The first sub-model computes the probability $P_n(R_i \mid h_j, c_j)$, by exploiting the probabilities of resource sequences of the *n*-grams model. The second sub-model is the distance-dependent trigger, it computes the probability $P_t(R_i \mid h_j, c_j)$ This last sub-model is devoted to resources out of the training corpus. A probability *a priori* $P_a(R_i \mid c_j)$ is set to each resource $Ri \in R$.

The resulting model, that can be viewed as a community-based Statistical Grammar of Usage, computes the linear combination of the three previously described sub-models.

$$P(R_i \mid h_j, c_j) = \lambda_n P_n(R_i \mid h_j, c_j) + \lambda_t P_t(R_i \mid h_j, c_j) + \lambda_a P_a(R_i \mid c_j) \quad (8)$$

Where λ_n, λ_t and λ_a sum up to 1 and are optimized with EM algorithm on a development corpus.

Thus, given a user U_j and his/her history h_j, we first have to determine the community c_j he/she belongs to. The probability of any available resource is computed given the SGU learned for this community. The N most likely resources are selected.

5 Conclusion and perspectives

This paper aims at describing a new web browsing assistant, based on usage and natural language processing. This approach exempts the difficult task of content, structure or format indexing and facilitates heterogeneous resources management. Similarities between SLM and web browsing are put forward, therefore the integration of usual statistical models from statistical language modeling domain is investigated. The resulting model is a Statistical Grammar of Usage (SGU). As a single SGU may be inefficient, it has to be personalized. To tackle sparsity of data, a preclassification of users into communities is performed. Community-based SGU are then proposed. Moreover, a new model is introduced, managing variable distance dependent triggers.

This new model is a first contribution to increase quality of prediction of resources in web browsing. A second contribution consists in the design of community-based SGU, predicting the sequentiality of resources during navigation. Moreover, a community-based SGU builds an *a posteriori* structure of navigation based on the subjective but reliable measure of pertinence of a resource for a user. Consequently it performs a personalized indexing of resources, based on usage analysis.

Collaborative filtering techniques used to build communities and triggers used to suggest resources have both proved their efficiency in their respective domain. A first perspective is the validation of the community-based SGU in terms of quality of predictions in web browsing. This evaluation can be performed by measuring the perplexity of the model. A second perspective is the use of the community-based SGU to compute a personalized classification of resources, depending not only on topics but also on user's preferences and context.

REFERENCES

Abramson, N. (1963). *Information Theory and Coding*. McGraw-Hill, New-York.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. ACM Press, New York.

Brun, A., Smaïli, K., and Haton, J. (2002). Contribution too topic identification by using word similarity. In *ICSLP2002*.

Castagnos, S. and Boyer, A. (2006a). A client/server user-based collaborative filtering algorithm model and implementation. In *Proceedings of ECAI 2006*, Italy.

Castagnos, S. and Boyer, A. (2006b). Frac+: A distributed collaborative filtering model for client/server architectures. In *WEBIST 2006*, Portugal.

Chan, P. (1999). A non-invasive learning approach to building web user profiles. In *KDD 1999 - Workshop on Web Usage Analysis and User Profiling*, USA.

Herlocker, J., Konstan, J., Terveen, L., and Riedl, J. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53.

Jelinek, F. and Mercer, R. (1980). Interpolated estimation of markov source parameters from sparse data. In *Wk. on Pattern Recognition in Practice*, pages 381–397.

Jurawski, D. and Martin, J. H. (2000). *Speech and Language Processing: an Introduction to Natural Language Processing*. Prentice-Hall.

Rosenfeld, R. (1996). A maximum entropy approach to adaptative statistical language modeling. *Computer Speech and Language*, 10:187–228.

Rosenfeld, R. (2000). Two decades of statistical language modeling: Where do we go from here.

Smaïli, K., Brun, A., Zitouni, I., and Haton, J. (1999). Automatic and manual clustering for large vocabulary speech re cognition: A comparative study. In *Eurospeech'99*, Hungary.