

# Association-Based Identification of Internet Users Interests

M. Charnine,<sup>1</sup> A. Petrov,<sup>2</sup> and I. Kuznetsov<sup>1</sup>

<sup>1</sup>Institute for Informatics Problems, Russian Academy of Sciences, Moscow, Russia

<sup>2</sup>Tinkoff Digital, Moscow, Russia

**Abstract** – *The method to provide the Internet user with useful information, while using search engines is presented. Here we mean the systematization of search results according to user's interests, and also showing advertisement which could be interesting for the user. We introduce concept of "user profile", consisting of keywords/terms, reflecting user interests. The discovering of such keywords is done by parsing user queries and visited websites. The proposed method uses a tree of categories linked to related websites and to the advertising. From these websites we retrieve primary keywords characterizing categories. The primary keywords are extended with new associated ones (called secondary) which were obtained by the methods of distributive semantics. By comparing keywords of user's profile and keywords of the categories we determine relevant categories and useful information (including advertising). The use of distributional semantics methods allows us to obtain good results even on short search queries.*

**Keywords:** user behavior analysis, user interests; cognitive technologies; semantic analysis of Internet texts; text classification.

## 1 Introduction

One way of increasing the quality of Internet search engines and other Internet systems is the use of knowledge about the characteristics of users. Such characteristics may be gender, age, interests of users in the form of keywords, etc. Having more complete information about the user, the search engines can provide more "personalized" results, as follows:

- Show to user the content which he/she is interested in (articles, music, movies, books, etc.);
- Adjust the sequence of the search results for the user. For example, the search engine can show more interesting to the user web page in a higher position.
- Optimize the selection of advertising, showing to the user only those web pages and pictures that can be interesting to him/her.

Typically, data about the features and characteristics of a particular user are not explicit in the Internet. At the same time, the Internet has a large set of data about the user's

actions while using the computer. By using such data it is possible to recover the missing information about users with sufficient accuracy. The data about the user actions includes:

- history of visits of the various Internet pages;
- user search history;
- history of clicks on hyperlinks and banners;
- history of purchases from online stores;
- video viewing history;
- history of listening audio files.

In this article, to determine the characteristics of a user we consider only two types of actions: the history of visits and the search history. We introduce a method that gives the possibility to automatically find the characteristics of users by linguistic analysis of search queries and visited web pages.

## 2 Representation of user interests

To solve the problem of automatic identification of user interests we chose an approach based on automatic analysis of the history of user queries and visited websites. We also use a training set of categorized webpages to build classifier that contains not only keywords/terms of the training set, but also associatively related terms. The analysis of the user history includes automatic classification of visited webpages and discovering/mining keywords/terms that characterize the current user.

The user's interests can be represented as a set of pairs <keyword, weight>. For example, a shopper can have the following interests: <dress, 70> <cosmetics, 40> <handbag, 20>, and the interests of the sports fan can be represented as follows: <sports, 30> <football, 60> <volleyball, 20>. Here the numbers are representing the weight of the keyword, which expresses a particular interest in a certain scale. Note that a specific user interest may also be represented as a point (or vector) in a multidimensional vector space of keywords/terms. This way a vector of user interests is formed. This representation allows us to introduce a measure of semantic similarity between interests of different users (proximity of points in keyword/term space). Also this representation helps us to automatically identify the groups of users with similar interests.

To classify the interests of users we use the category tree (taken from the Internet), where each category corresponds to a specific theme, expressed by the keywords or terms (football, indoor football, American football) and has links to the relevant explanatory texts (websites, articles). These texts are used as a training set for the process of automatic categorization and discovering keywords/terms that are related to each category. These category keywords/terms are represented as triples  $\langle \text{keyword}, \text{category}, \text{weight} \rangle$ , where weight is a number that describes the significance of the keyword for identification of the category. Let's call such keywords as primary. The number of such triples can be extended by including of keywords/terms that are associated with the name of category or with a set of primary keywords of the category. These associated keywords/terms are discovering from the variety of Internet texts where the name of the category, for example, football, is mentioned. Let's call such associated keywords/terms as secondary keywords. This way we get triples  $\langle \text{secondary keyword}, \text{category}, \text{weight} \rangle$  where the weight describes the strength of the association.

In proposed method each category in the category tree is associated with relevant content or advertising. So it's very important to determine the appropriate categories for the user. This determination is done by comparison of the keywords of the user and the category. The categories, to which the user is belong/relevant, are chosen because of their semantic similarity. As a result, the user receives relevant content or advertising.

Note that if we know the vector of user interests, then it is possible to make an assumption about the gender and age of the person. For example, men are more interested in football, and women are more interested in cosmetics. Young people are interested in contemporary music and at the same time old people are interested in a retro. We call vectors of user interests, which are supplemented by information on the categories of user, the user's profile.

### 3 Building of the vector of user interests

The calculation of vector of user interests is done by automatic analysis of the history of user queries and visited websites. This user history determines the collection of user related texts including texts of queries and texts of visited web pages.

Vector of user interests is constructed by statistical analysis of user related texts and calculation of frequencies of the various keywords/terms that are mentioned in these texts. These data forms a multidimensional vector of user interests, the components of which are the frequencies of the various keywords. This vector also can be represented as a point in a multidimensional vector space of keywords/terms.

A set of primary and secondary keywords of each category and their weights also can be presented as a point (or vector) in this multidimensional vector space. This vector representing weights of primary and secondary keywords of category we called as vector of category features.

This representation allows us to introduce a measure of semantic similarity between vector of user interests and vector of category features as proximity of points in keyword/term space. So this representation helps us to automatically identify the most relevant categories for the user.

Vector of category features depends from the method of calculation of weights of primary and secondary keywords. So weights of primary and secondary keywords of category must be carefully calculated to provide the most accurate classification of text documents and users histories.

### 4 Calculation of the optimal weights

The task of automatic classification of visited documents (their relation to the categories), can be reduced to the question of collective decision of experts, each of them only responds to the document in the presence of one particular keyword/term associated with the category. The experts have to answer the question: is the document belongs to category or not? Thus, the number of experts is equal to the number of keywords and each expert is responsible for one specific keyword. If the keyword exists in the document, the expert makes decision about belonging of the document to the specified category. The probability of correct expert decision is equal to conditional probability:

$$P_i = p(\text{Category} | \text{Keyword}).$$

The main result of Nitzan and Paroush (1982) [16], and Shapley and Grofman (1984) [17] was the claim that if we have two alternatives and the probability of correct decision for each expert is known, then the linear combination rule of their decisions is optimal, and the maximum probability of correct collective decision is reached when the weights (shares) of experts calculated by the formula:

$$W_i = \log(P_i / (1 - P_i)), \quad (1)$$

Where,  $P_i$  – is the probability of correct solutions of expert with number “i”. This formula (1) is optimal when we have two alternatives, for example, gender: man and women.

In the case when we have more than two alternatives we can use another formula (2) that is discussed below.

Let Keyword represent the problem keyword and {category-1, category-2, ..., category-n} represent the alternatives. The PMI-IR algorithm assigns a weight/score to each category,  $\text{weight}(\text{category-i})$ , and selects the category that maximizes the weight/score.

The PMI-IR algorithm, is based on co-occurrence. The core idea is that “a word is characterized by the company it keeps” [21], that is another version of distributional hypothesis discussed later. There are many different measures of the degree to which two facts co-occur. PMI-IR uses Pointwise Mutual Information (PMI) [19, 20], as follows:

$$\text{weight}(\text{keyword}) = \text{weight}(\text{category-i}) =$$

$$\text{Log}(p(\text{keyword}\&\text{category-i})/(p(\text{keyword})p(\text{category-i}))) \quad (2)$$

Here,  $p(\text{keyword}\&\text{category-i})$  is the probability of co-occurrence of two facts: keyword exist in the document and document belongs to category-i. If keyword and category-i are statistically independent, then the probability that they co-occur is given by the product  $p(\text{keyword})p(\text{category-i})$ . If they are not independent, and they have a tendency to co-occur, then  $p(\text{keyword}\&\text{category-i})$  will be greater than  $p(\text{keyword})p(\text{category-i})$ . Therefore the ratio between  $p(\text{keyword}\&\text{category-i})$  and  $p(\text{keyword})p(\text{category-i})$  is a measure of the degree of statistical dependence between keyword and category-i. The Log of this ratio is the amount of information that we acquire about the presence of keyword when we observe document of category-i. Since the equation is symmetrical, it is also the amount of information that we acquire about belonging the document to the category-i when we observe keyword, which explains the term mutual information.

These formulas (1) and (2) allow us to calculate weight of keyword associated with a given category if we know the probability of the presence of this keyword in the documents of the category. The method of calculation of probabilities is discussed in the next section.

## 5 Calculation of the probability of finding given keyword in the context of the category

Let's define the context of the category as a set of documents (or sentences) associated with this category. In the simplest case, the probability of presence of given keyword/term in the context of the category is calculated as the number of occurrences of this keyword in the context divided by the number of occurrences of keyword in all analyzed texts.

This method is similar to the wellknown measure TF-IDF [5], which takes into account not only the frequency of the presence of keyword/term in the documents associated with the category, but also the frequency of documents containing given keyword/term, which reduces weight of generally used and insignificant words such as interjections, prepositions, etc.

A more accurate calculation of the probability of occurrence of the keyword/term in the context of the category should consider the probability error that is discussed in section 9.

## 6 Calculation of the weights of primary keywords of the category

To calculate the weight of primary keyword of the category we should know the probability of presence of this keyword in the context of category. We use category tree with related documents to calculate the context of category.

As a training set for the creation of the initial version of the category tree we can use any online directory that contain links to Web sites, for example, Google, Yahoo, Yandex. In our research we use Yandex Catalogue as a reference catalogue of Russian sites. Yandex Catalogue consists of a set of tree-like categories; each of category contains a brief description, subcategories and links to Web pages. Thus one web page can be included into several different categories. In this case, if the web page belongs to the category and it also belongs to all parent categories.

For example, a category FOOTBALL (upper level - SPORT) consists of subcategories:

RUSSIAN FOOTBALL

WORLD FOOTBALL

MINI FOOTBALL. . . ,

And links to sites:

FOOTBALL ON PORTAL "CHEMPIONAT.COM"

Russian Football Union. . .

Websites have supplementary texts and pictures which are used as a reference to the sites.

We use the program in python and library nltk to collect text information from web pages. Text of web pages has been cleared of html-markup and broken into strings of words by using standard library functions “nltk”. We divided such strings into various keywords. For each keyword we have counted the number of occurrences in the document. As a result, for each web page from the referenced file directory has been created a file, which consists of triplets

<keyword, category, the number of occurrences>.

On the next step with the data processing program map-reduce [6] for each pair <keyword, category> was counted the total number of occurrences of the keywords in each of the parent categories. The resulting file was used to assign keywords to a category.

These data were used for calculation of the probabilities of occurrence of keywords in the documents of particular category. By using these probabilities and formulas (1) and (2) we calculated the weight of primary keyword of the category.

## 7 Calculation of the weights of secondary keywords of the category

The weight of a secondary keyword, which is associated with the name of the category, is calculated using the probability of occurrence of the keyword in the context of the category. Context is the set of all documents (or phrases) containing the name of the category and some of its primary keywords.

Discovering of the association from Internet texts is based on the distributional hypothesis, which states that semantically similar (or related) keywords/terms have a similar context, and, conversely, keywords/terms with similar context are semantically close.

The discovery of secondary keywords and their weights involves the use of various methods, including:

- methods of detection of Internet texts of specific subject areas;
- methods of detection of significant texts keywords, terms, and their ranking;
- methods of identifying and ranking the associative relationships between important keywords and terms.

The processing of large volumes of texts that are constantly updated in the Internet allows us to collect all the necessary statistical data to generate a fairly complete picture of the subject area that can be represented as a set of associative relationships. The ability to use machine learning techniques on a large number of examples gives the system flexibility and improves the results.

## 8 Aligning of weights of primary and secondary keywords

The weight of the keyword that is associated with the category may be calculated by different methods, for example, based on the probability of occurrence of the keyword in the documents of category in training set. The next possible method for calculation of the weight is based on the probability of finding keyword in the Internet context of the category. Saying context here we mean all the Internet documents (or phrases) that contains category name and some of the primary keywords of the category.

The weights of the same keyword calculated by the different methods should be similar as much as possible. To achieve that, we proposed the following method of weights aligning.

Suppose, by using two different methods, we have found two groups of keywords T1 and T2 that are associated with a particular category. For example, a set T1 consists of primary keywords that are included in the training set of the category, and a set T2 consists of secondary keywords associated with

the category and discovered from Internet context of the category. The Internet context of the category usually is much bigger than training set, so the set T2 is usually bigger than T1. The weights for T1 and T2 are calculated independently by using two different algorithms. Usually the set T1 is included into the set T2, so the weights for T1 are calculated by both algorithms, and we can select the weight coefficient for T2 to align them with the weights of T1. We use this coefficient for calculating all the weights of secondary keywords.

## 9 Determining the user's interests using history of his visits and inquiries

Previously, it was discussed the method of calculation of the category of each of the visited by user web pages. This method can serve as the basis for construction a user profile that will reflect his interests. Such profile can be represented as a plurality of pairs <category, weight>. Here the weight can be calculated, for example, as the number of the web pages in the category «C» divided to the total number of web pages, visited by the user.

Suppose that the user's interests are unchanged. Then we can assume that the number of web pages of the category «C» in comparison with the total number of visited web pages obeys the binomial distribution. Let's take «N» as the total number of pages that were visited by the user, and «n» - the number of pages category «C». In this case, the weight will be calculated according to the formula:  $W = n/N$ . Then confidence interval (with 95% confidence level) may be calculated by the formulas:

$$W_{min} = W - 1.96 * \text{SQRT}(W*(1-W)/N).$$

$$W_{max} = W + 1.96 * \text{SQRT}(G2*(1-G2)/D2)$$

For example, if  $n = 2$  and  $N = 4$ , then in this case

$W_{min} = 0.01$ , and  $W_{max} = 0.99$ , then there we can say nothing specific about the value of  $W$ . If  $n = 100$  and

$N = 200$ ,  $W_{min} = 0.43$ , and  $W_{max} = 0.56$  - in this case, it is safe to say that the category is found and that it is not a "noise" (i.e., web pages that were visited by accident). In order to exclude the "noise" from the user's interests we included into the user profile only those keywords which have the minimum value of the confidence interval greater than a certain threshold.

## 10 Conclusions

This paper describes the method of determination of user's interests from the history of visited websites and user search history. We have developed a program that can successfully determine the interests of the user even from the short texts that do not contain keywords/terms from training set for calculation of the classifier. Note that standard algorithms usually didn't work well with such texts. In the future we plan

to develop an algorithm that will allow us to get the user profile, which takes into account a number of factors, including the user gender, age, interests, intentions, region of residence, income level, marital status and other useful information.

## 11 Acknowledgements

This work was supported by the Russian Foundation for Basic Research, grant #13-07-00272 “The methods for automatic creation of associative portraits of subject domains on the basis of big natural language texts for knowledge extraction systems”.

## 12 References

- [1] Greg Linden, Brent Smith, Jeremy York “Amazon.com recommendations. Item-to-item collaborative filtering”. EEE Internet Computing, Los Alamitos, CA USA, 2003 <http://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf>.
- [2] Brian McFee, Luke Barrington, Gert Lanckriet “Learning Similarity from Collaborative Filters”, ISMIR 2010, [http://cosmal.ucsd.edu/cal/pubs/ISMIR2010\\_learnCF.pdf](http://cosmal.ucsd.edu/cal/pubs/ISMIR2010_learnCF.pdf).
- [3] M.C. Agueyev, "Methods of automatic text categorization based on machine learning and expert knowledge", PhD Thesis: 05.13.11, Moscow, 2004.
- [4] V.E. Abramov, "Automatic classification and abstracting of textual information: including in foreign languages," PhD Thesis: 05.25.05, Moscow, 2008.
- [5] Salton, G. and McGill, M. J. 1983 Introduction to modern information retrieval. McGraw-Hill, ISBN 0-07-054484-0.
- [6] J Dean, S Ghemawat, «MapReduce: simplified data processing on large clusters», Communications of the ACM, 2008 - [dl.acm.org](http://dl.acm.org).
- [7] E.Baharad, J.Golberger, M.Koppel и S.Nitzan, “Beyond Condorcet: Optimal Aggregation Rules Using Voting Records”, CESifo München, 2011.
- [8] A. Lenci, “Distributional semantics in linguistic and cognitive research”, *Rivista di Linguistica*, 1, 2008, pp.1-30.
- [9] M.Baroni, A.Lenci, “Distributional Memory: A General Framework for Corpus-Based Semantics”, *Computational Linguistics*. V.36, Issue 4, 2010, pp. 673-721.
- [10] Peter Turney, “A uniform approach to analogies, synonyms, antonyms and associations”, *Proceedings of COLING*, Manchester, 2008, pp. 905–912.
- [11] Peter Turney, “Mining the web for synonyms: PMI-IR versus LSA on TOEFL”, *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, Freiburg, Germany. September 3–7, 2001. pp. 491–502.
- [12] M.Charnine, I.P.Kuznetsov, E.B.Kozerenko, “Semantic Navigator for Internet Search”, *Proceeding of International Conference on Machine Learning*, 27-30, 2005 Las Vegas, USA, CSREA Press, pp. 60-65, 2005.
- [13] Michael Charnine, Vladimir Charnine. *Keywen Category Structure*.// Wordclay, USA, 2008, pp.1-60.
- [14] Michael Charnine, “Keywen Automated Writing Tools”, Booktango, USA, 2013, ISBN 978-1-46892-205-9.
- [15] Nitzan, S., and J. Paroush. 1985. “Collective Decision Making: An Economic Outlook”, Cambridge University Press, Cambridge, England.
- [16] Nitzan, S., and J. Paroush. 1982. “Optimal Decision Rules in Uncertain Dichotomous Choice Situations.” *International Economic Review* 23(2): 289-97.
- [17] Shapley, L. and B. Grofman. 1984. "Optimizing Group Judgmental Accuracy in the Presence of Interdependencies." *Public Choice* 43: 329-343.
- [18] R. Diamond, “Designing and improving courses and curricula in higher education: A systematic approach: Jossey-Bass,” SF,1989.
- [19] K.W. Church, P. Hanks, “Word association norms, mutual information and lexicography”, *Proceedings of the 27th Annual Conference of the Association of Computational Linguistics*, 1989, pp. 76-83.
- [20] K.W. Church, W. Gale, P. Hanks, D. Hindle, “Using statistics in lexical analysis”, In: Uri Zernik (ed.), “Lexical acquisition: Exploiting on-line resources to build a lexicon”, New Jersey, Lawrence Erlbaum, 1991, 115-164.
- [21] Firth, J.R.: “A synopsis of linguistic theory 1930-1955”, In *Studies in Linguistic Analysis*, pp.1-32, Oxford, Philological Society (1957).