InfoCodex
Semantic Technologies

Bahnhofstrasse 50
CH-9470 Buchs
www.infocodex.com

Tel: +41 (0)81 750 53 00
Fax: +41 (0)81 750 53 01
info@infocodex.com

23.07.2008

Ads Matching in Online Advertising – A Turnkey InfoCodex Solution

1. Scenario

From a large pool of advertisements, the ones that best match the content of the active Web page (or the personal interests of the user) should be selected. The mapping can either be based on:

- The content of the active Web page ("contextual targeting") or
- The contextual analysis of the Web pages recently visited by the user and therefore taking into account the user's preferences ("behavioural targeting").

The impact of advertising critically depends on the quality of the mapping, i.e. on how well the advertisement matches the user's current interests. This is the case for both mapping strategies.

The diversity of Web pages and available advertisements is enormous, and therefore a matching simply based on the matching of keywords is rather limited. This is further exasperated if multiple languages have to be considered.

2. Solution

Starting position

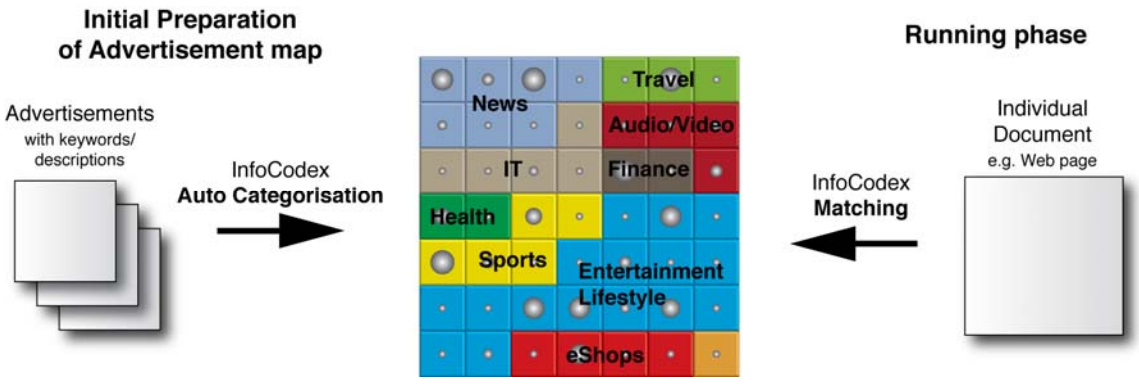
A large number of advertisements (e.g. 1000 up to several millions) are available, each with characteristic keywords or a short textual description (in English, German, French, Italian, or Spanish). Optionally, the advertisements may be assigned to given advertising categories. In this case, the individual advertisements don't require keywords or descriptions if the individual advertising categories are characterised appropriately.

Goal

For a given document (Web page, RSS feed, etc; in D, E, F, I or ES) the advertising categories or even individual advertisements should be selected which best match the content of the document.

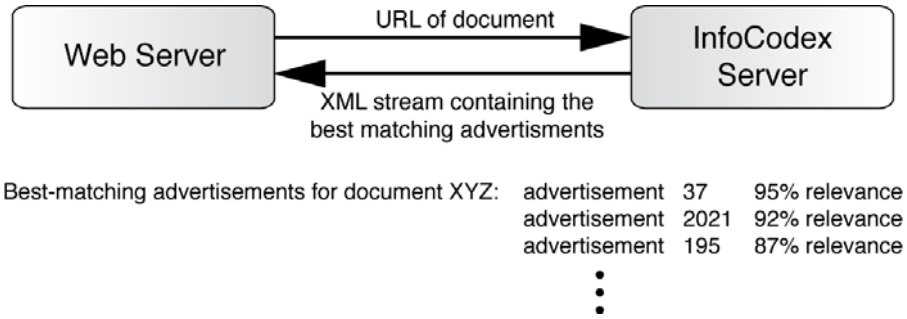
Solution with InfoCodex

During an initial preparation phase, the available advertisements are analysed for their content through analysing the associated keywords or textual descriptions. Using this information, the advertisements are then categorised into a structured "advertisement map" (virtual bookshelf). Advertisements with similar content are allocated to the same category, i.e. a particular "compartment" of the virtual bookshelf contains advertisements of similar contents. The categorisation is carried out fully automatically by InfoCodex and without human intervention. If desired, a user-specified advertising-taxonomy can be stipulated.



In the following running phase, the InfoCodex system extracts the *real content* from the documents (for which best-matching advertising categories or even individual best-matching advertisements are sought) by eliminating navigation elements, select boxes, links etc.). Then, the documents are analysed for their content and placed into the prepared advertisement map using a well-founded *content similarity measure*.

The result returned by the InfoCodex server for each requested document-categorisation consists of a short list containing the most relevant advertisements and their respective relevance.



The matching procedure is truly cross-lingual and takes into account the *effective content* of the categorised document (Web page, RSS feed etc), i.e. it produces good results even in cases where a simple matching based on keywords is not effective or if the considered documents are written in different languages. An English and a German web-page with an equivalent content are recognised by the system as very similar documents.

The matching mechanisms are discussed in more detail below.

3. Recognition of the content of a document by InfoCodex

Step 1: Content extraction

Prior to the effective content analysis a filter is applied that extracts the *real content* of a Web page by removing navigation elements, select boxes, links, and advertisements. The content extraction works fully automatic. It can optionally be controlled by the following parameters

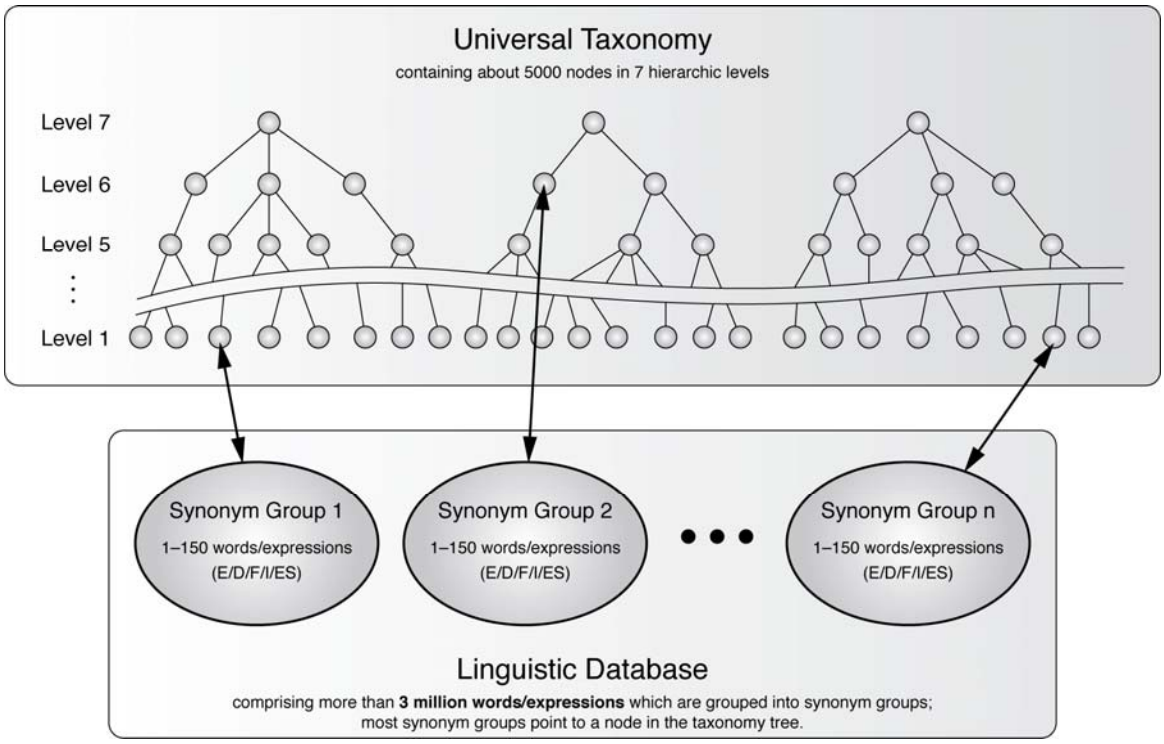
- Minimum size of the text blocks that are to be extracted (default: 20 words without counting the tags in the text block)
- Minimum size of the text blocks to be extracted if none of the recognized blocks exceeds the above-mentioned limit (e.g. in the case where the HTML page consists of an image and a very short description)
- List of tags that do *not* interrupt a text block (e.g.
, <b>, etc.)
- List of tags that terminate a text block
- List of tags that have to be removed or must be included in any case, respectively.

Step 2: Content recognition and similarity analysis

Foundation: multi-lingual database linked to universal taxonomy

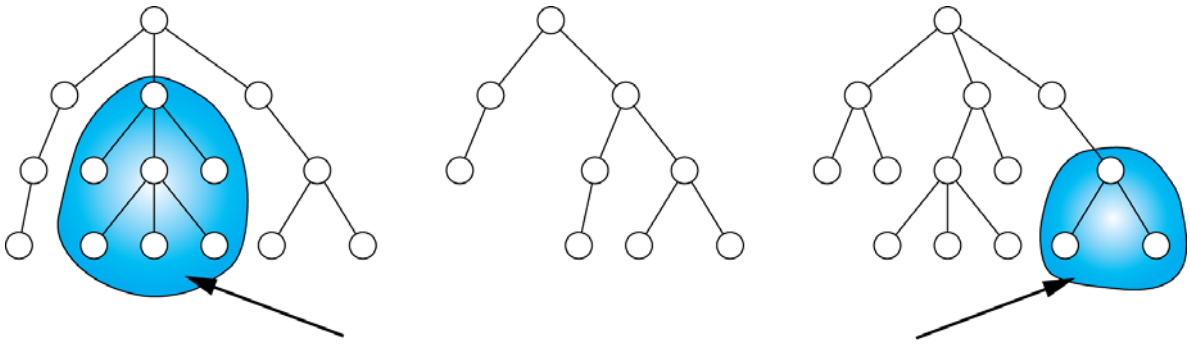
In the context of InfoCodex' linguistic database the term "taxonomy" refers to a *taxonomy (simplified ontology) for single words/expressions*. In contrast, the term "advertising taxonomy" usually refers to a categorisation of advertisements (*application-specific table of contents for entire documents/articles*).

InfoCodex' linguistic database comprises more than 3 million words and expressions (groups of words such as "European Union", "Enterprise Search Engine" etc.) in currently five languages (E/G/F/I/ES). These are grouped into synonym groups which are then systematically linked to a universal taxonomy.



Content analysis of an individual document

The words or expressions present in the document are matched against the linguistic database and their meaning is determined by the links of the corresponding synonym group of the linguistic database to the universal taxonomy. During the analysis of an individual document, all nodes in the taxonomy tree which are addressed by the matching process are highlighted, and the ensemble of highlighted nodes is a measure of the thematic areas covered by the document.



Highlighted taxonomy tree nodes for an individual document

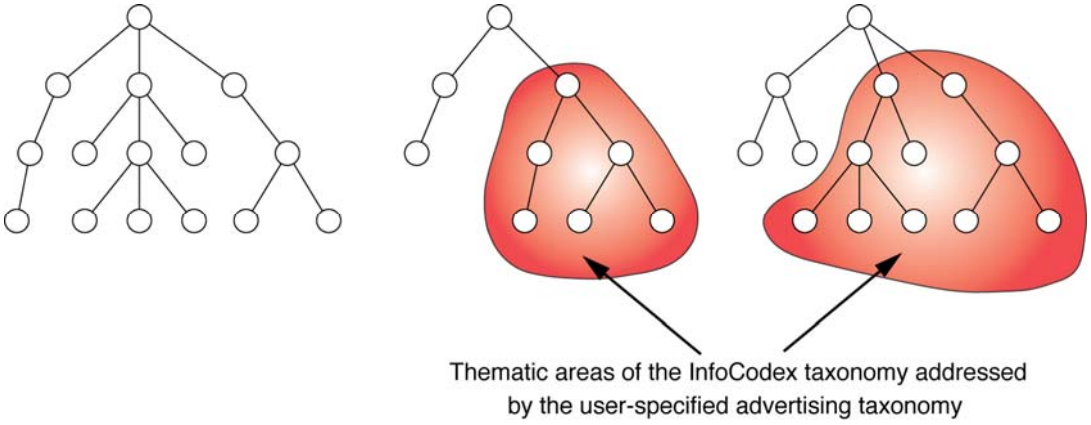
The addressed areas of each document are then projected into a **100-dimensional content-space**, and finally a categorisation of the documents is achieved by means of a **self-organising neural network** (Kohonen-Map). The categorisation of a document by InfoCodex is not a simple assignment to a single node in the taxonomy tree, but rather a multidimensional projection.

This neural network provides a **well-founded similarity measure** based on information-theoretical principals which allows the comparison of documents according to their content. A small distance between documents corresponds to a high similarity and vice versa. In mathematical terms, the similarity measure is given by the weighted scalar product of the two vectors, corrected by the Kullback-Leibler distance from the main themes, combined with the weighted score-sum of the matching keywords and their nodes in the taxonomy tree, respectively.

The similarity measure is independent of the language of the document and is only weakly dependent on the exact wording. The described processes are patented in the EU and USA.

4. Benefits of superposition of user-defined advertising taxonomy

The user can stipulate an advertising taxonomy (fixed advertising categories), which is tailored to the specific requirements of the user, i.e. is focussing only on advertising-relevant thematic areas. Such a taxonomy is *supplied as a simple Excel table* that comprises the hierarchic advertisement categorisation scheme and some optional descriptions characterizing the individual advertisement categories. It covers the advertising-relevant part of the knowledge spectrum.

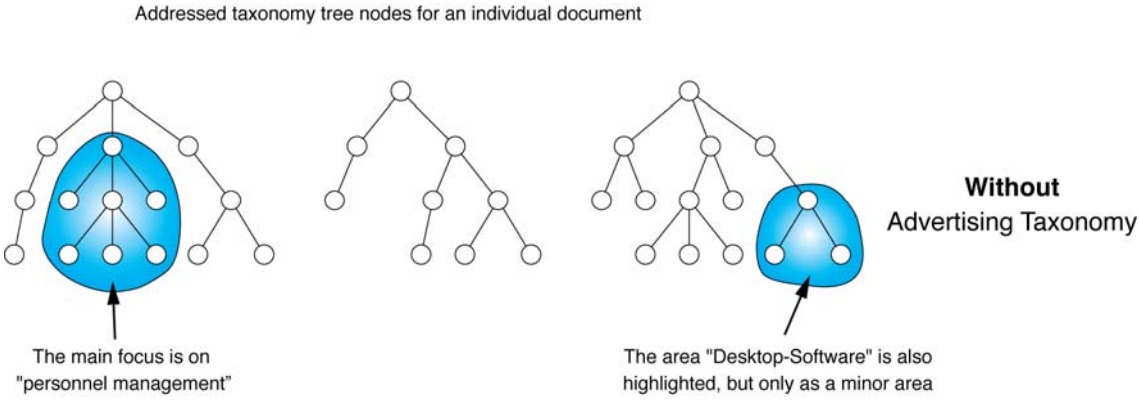


The advertisement taxonomy acts as a pair of glasses which put the focus onto advertisement-related themes and suppress everything else.

Example illustrating the mechanism of an advertising taxonomy

Assume in a given document (Web page, RSS feed, etc) primarily “personnel management” is discussed. However, the main theme “personnel management” is not one of the nodes in the advertising taxonomy. The only nodes of the advertising taxonomy that bear limited relevance to the document are “Desktop Software”, and to an even lesser extent, “Internet games”.

a) Self-generated categorisation scheme

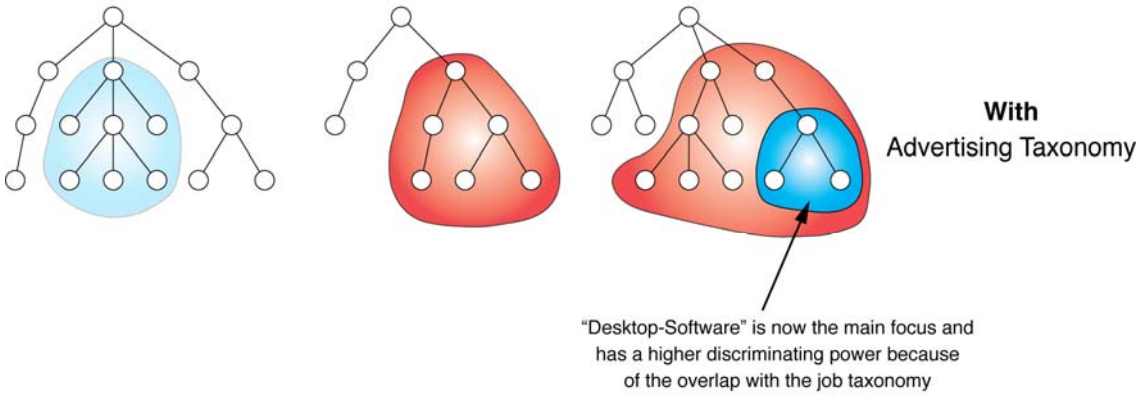


In a fully automatically built classification scheme, topics like “personnel management” and other non-advertising-related sections of the knowledge spectrum will be treated in the same way as advertising-related sections, and hence will compromise and reduce the discriminating power of the advertisement-related topics for the content analysis.

b) User-specified advertising taxonomy

Through the superposition of the advertising taxonomy, no advertising-relevant information is lost. Rather, only the advertising-relevant information is used for the categorisation, while the remaining, not relevant, information is suppressed. This leads to an improvement of the matching quality.

Addressed taxonomy tree nodes for an individual document: only areas that overlap with the user-defined advertising taxonomy are taken into accout



With
Advertising Taxonomy

"Desktop-Software" is now the main focus and has a higher discriminating power because of the overlap with the job taxonomy