

Vis4



visualization for

from data to knowledge through the visualization

Vis4Search – Product Presentation

The requirement:

One of the main problems that afflict web navigators and, generally, anyone has to do with the information retrieval into blocks of heterogeneous, unstructured and often uncorrelated data is the requirement of tracing, in the shortest time, the intended information.

The present limits:

The most common means it generally devolves upon in order to ease the above mentioned requirement is the one we usually define “research engine”; to be more precise it is that system that, from a set of data (web pages, documents, mails, files, etc.) performs such indexing operations as to promptly and efficiently consent the access to the information, guaranteeing in the majority of cases enough satisfactory results. However, generally, these systems encounter an annoying problem: the data are index-linked on the basis of logics that do not reflect the semantic of the content, but are based on reasons and weight assignment that varies from one search engine to another and are usually related to advertising logics. Furthermore, if the indexer has performed on a substantial block of information, on average the search results are spread on numerous pages, discouraging the majority of users to carry out an extended check of the obtained result.

In simple terms, often it may happen not to find what it is searched for simply because the subject of interest is on the 30th page of search results and the person who is searching is therefore forced to establish a more limited set of keywords in the hope of being able to identify the right combination in order to find the subject of his interest.

The new solution: Vis4

It is precisely in order to resolve this problem, or at least to try to improve the user experience that Vis4Search was conceived.

In fact, Vis4Search is a software that, met between the search engine and the user, performs syntactic and semantic elaboration of the results obtained from the engine, in order to produce a graphic representation of the query, based on the grouping of results by concepts.

Although apparently it may seem complicated, in fact the basic principle is very simple: Vis4Search analyses for us all the snippets (that are small phrases that the search engine renders and which give us a brief description of the single found entity) and categorizes them (operation technically called clustering) on the base of "semantic proximity" (or meaning related) by grouping exactly those results that according to its logic have elements of common contents, into a single conceptual group. Once defined all these groups, it does nothing but return, to the user, a **navigable graphic representation**.

Let us clarify this concept with a graphical representation. Within Figure 1 we see the system applied to a generic search of the keyword Armstrong on a search engine (eg Google). At the bottom we see the presence of a snippet graph. This is achieved starting from the block of snippets rendered by the engine, by applying a first instance of the algorithm underlying the clustering. Technically, each node (green dot) of the graph represents a search result. Between two nodes there is an arc (black connecting line) if and only the two results share a set of words. The weight of this arc (that is the thickness of the line) varies according to how strong the bond between the two nodes is.

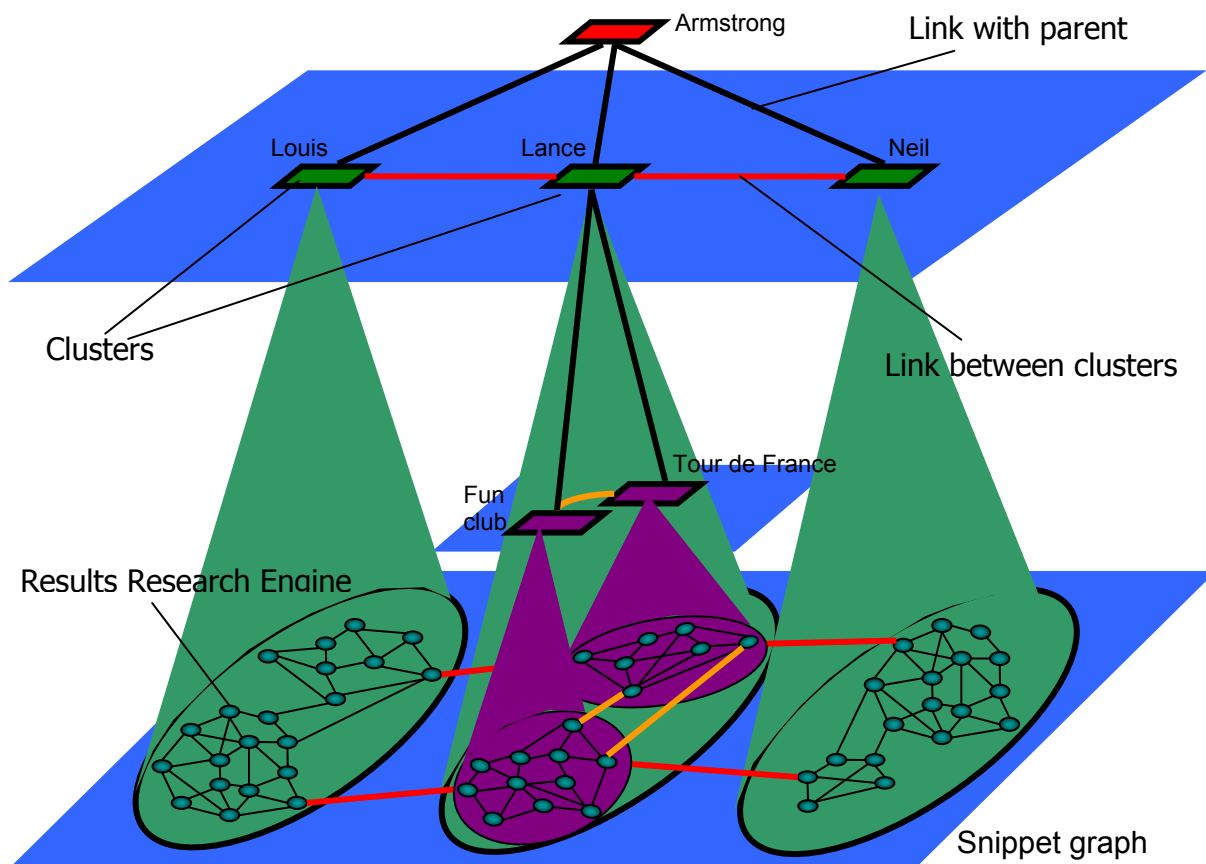


Figure 1 – Clustering system

Once obtained this graph, we will be able to define “dense zones” that are parts of the graph where more nodes are linked with each other (within the diagram highlighted by coloured cones). Each of these zones can be defined as a cluster. Clearly, if a cluster contains a large number of results, to these ones can be reapplied the algorithm by defining the sub-cluster and so on. To each cluster will be fasten a **tag** that is one or more words which, according to the algorithm, exhaustively represent the contents of the cluster. In order to stay on topic with the example, if one of the clusters contained a number of pages dedicated to the biography of Neil Armstrong, then, presumably, the algorithm would assign to cluster the tag Neil. Or even more if the cluster contained many references to the “tour de France”, then the cluster would be tagged

with the phrase "Tour de France", and so forth. The result would be, so, for the end user, a graphical representation of the first level clusters, with their relative tags, navigable that is to say such studied as to allow progressive access to information if there are sub-clusters.

Currently, the software, in terms of graphical representation, has two possible solutions:

- **Treemap:** is a representation in which each cluster is represented by a rectangle, whose size and position depend on how many results contain the square (size) and how much is related to the result with the content (location). In this representation each cluster is clickable, and clicking, a window will be displayed containing sub-clusters if present, otherwise the list of snippets. An example of a treemap representation is the Figure 2. We either note that in this representation, apart from the position, also the colour indicates how much the cluster is related to the research. Put simply, the clusters from the upper left with darker colour are more pertaining to the research, while the less relevant clusters are in the lower right with clearer colour. Generally, in this area the system poses a standard cluster, called "Other Topics", where are grouped all those results that are not semantically related in any way with the other results.
- **Radial:** this representation has a central element (the sought word) to which are linked by a line, all the first level clusters. The single clusters are placed in a circular shape and, pointedly, are clickable. At the click, the node will be expanded viewing either directly the individual pages that are part of the cluster, clicking on which the mentioned site is opened, or the sub-clusters in the same manner as the first level ones. Within the Figure 3 we see a representation of this type of display. The particularity of this graphic solution deals with the possibility to have, at a glance, an overview of the outcome dimension and its distribution. In this case also, the colour gives us an indicator of how pertaining is the content of the cluster: more is dark the tonality, more the content of the cluster is pertaining to the search key.

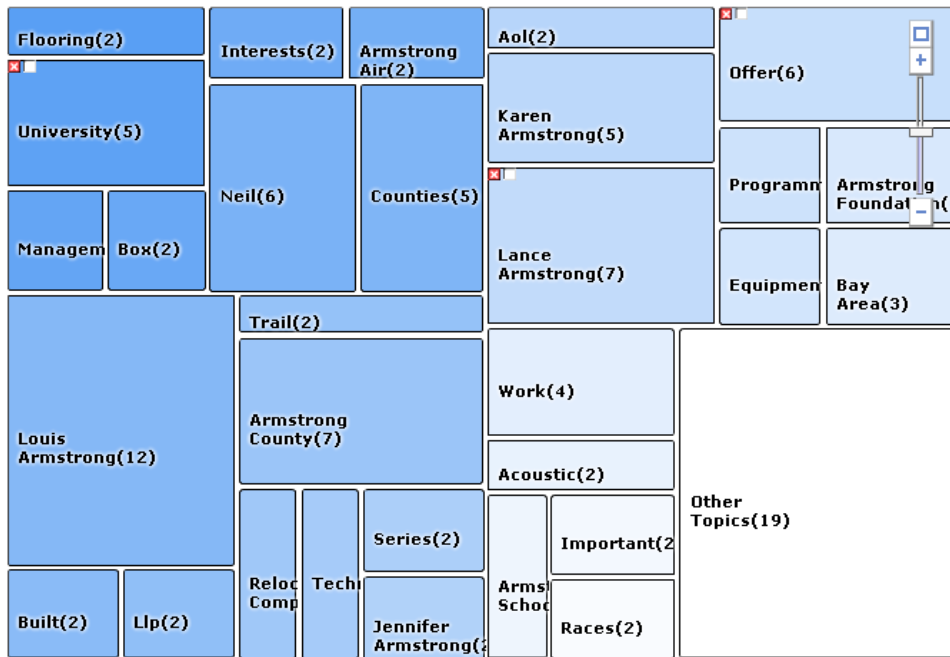


Figure 2 – Treemap Representation

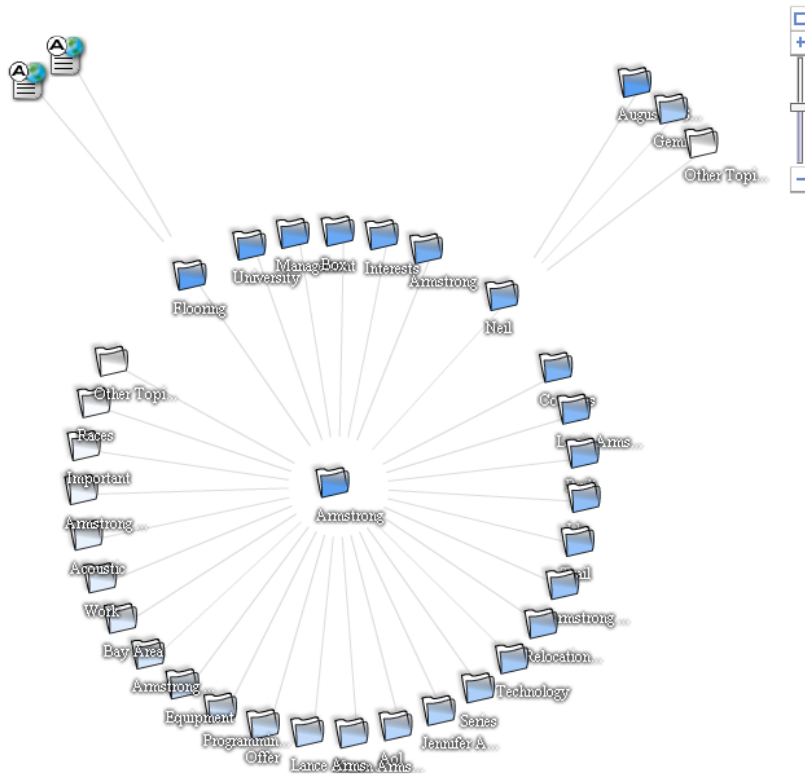


Figure 3 – Radial Representation

The other feature, available within the graphical representation and common to the two proposed types, regards the highlighting of the **correlations** between clusters. In other words, diverse clusters can be correlated with each other because between them there is a semantic connection (as seen within the Figure 1, a cluster is given by "dense areas" of nodes, but these dense areas can be linked to other dense areas because they semantically share some information. In these cases it is useful to highlight which are the correlated cluster because they can give us an useful indication of where can be contained other information related to the cluster of interest. Graphically this can be done by placing the mouse over a cluster for about two seconds, after which the system will highlight the related cluster hiding all other clusters in the case of the treemap representation (Figure 4) and explicitly highlighting connecting arches in the case of radial representation (Figure 5).

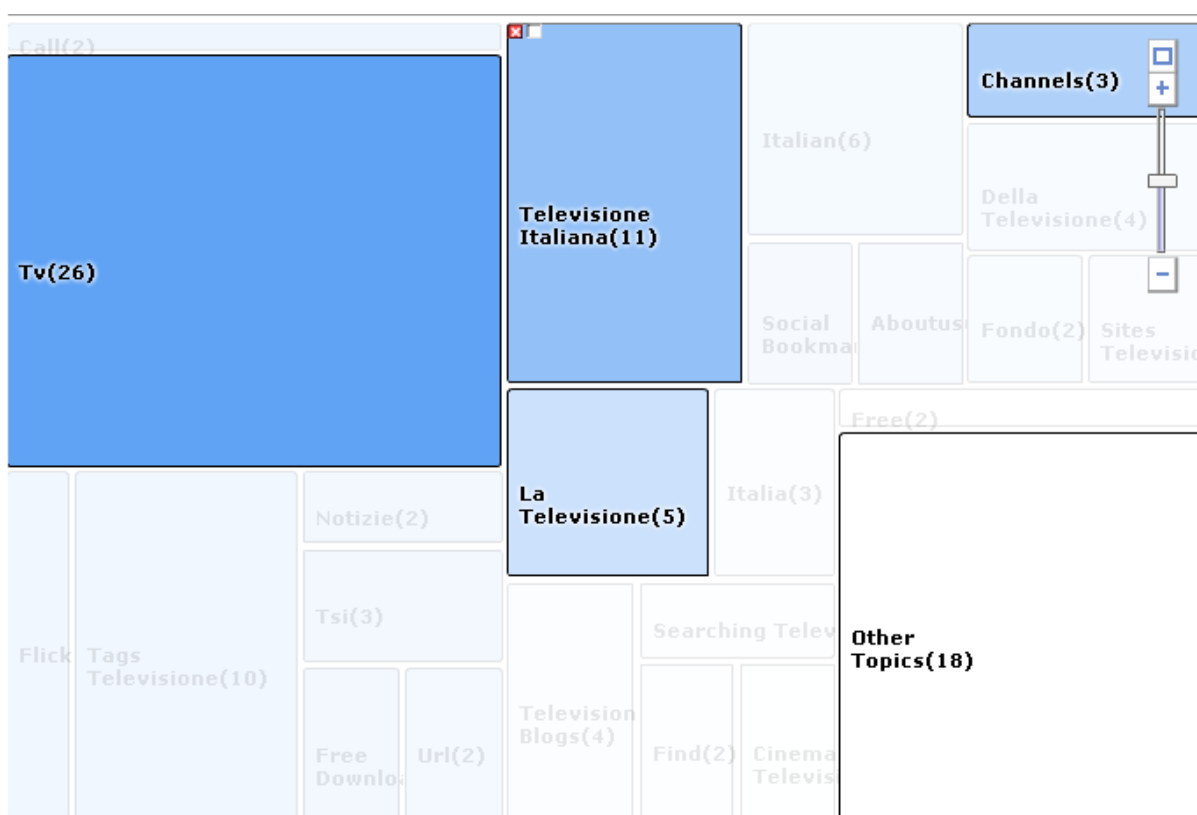


Figure 4 – Highlighting of clusters within the treemap representation.

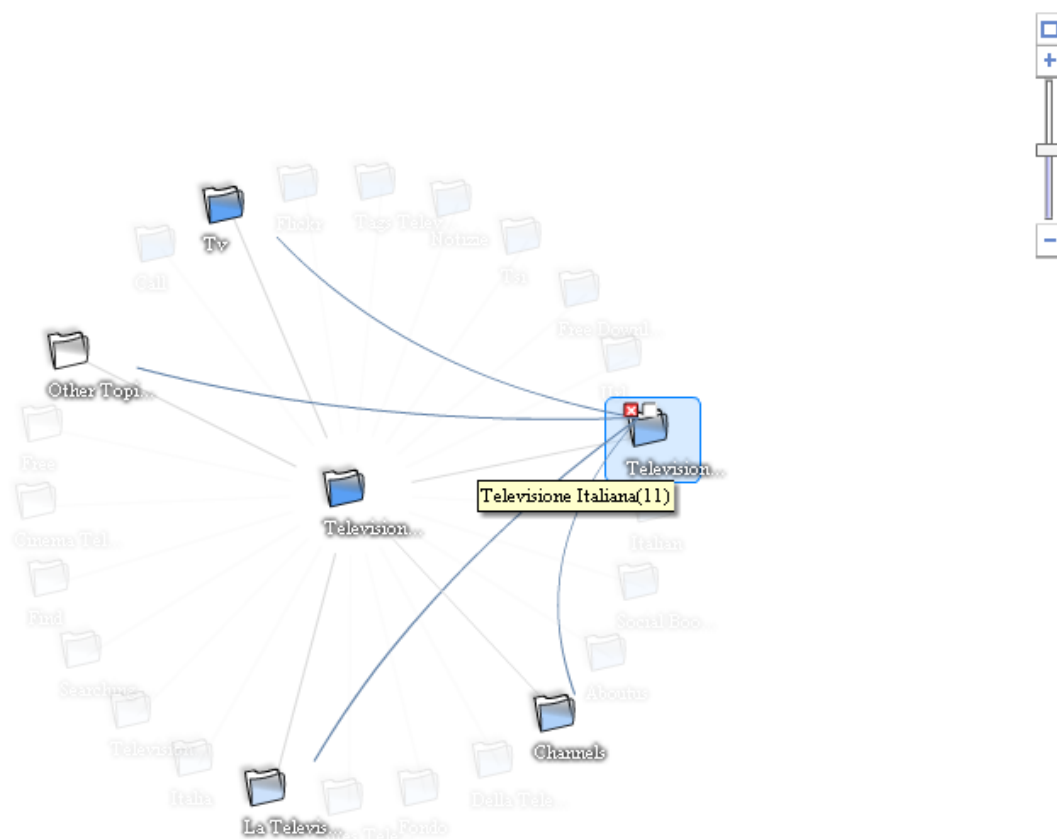


Figure 5 – Highlighting of clusters within the radial representation

Finally we highlight an additional functionality made available by the system. In particular, as we have already said, to each cluster is assigned a tag, which generally represents one or more keywords that exhaustively indicate the contents of the cluster itself. Having at our disposal this information, we can study a **dynamic advertising system** which, based on the tag of the currently visited cluster, presents a new set of banners chosen among those that have a correlation with the navigated argument. Let's make a practical example: from the query Armstrong previously described, in general we obtain several clusters related to the different known meaning of Armstrong. In particular, for example, we can simply restrict ourselves to Armstrong cyclist, musician and astronaut. Assuming that the system has clustered the three meanings in three different clusters, respectively tagged with, for example, "tour de france", "jazz" and "gemini", in case the user has decided to navigate the cluster "jazz", advertisement banners connected to the sale of musical instruments or concerts presentation can be loaded and displayed; if, instead, the choice is to browse the cluster "tour de france" banners relating to the resale of bicycles and sport magazines, specialized in cycling, are displayed and so on.

This would help to manage an advertising system updateable in an "intelligent" manner according to the choice made by the user through the navigation.

To sum up, such a system, developed at the **Electronic and Information Engineering Department** of the **University of Perugia**, proposes a viable solution in order to facilitate the navigation of our data. Obviously such a system, placed as an intermediary layer between the indexer (search engine) and the user, can be fastened, with minimal effort, to any existing indexing system, applied to any set of data. It is therefore possible to use the clustering functionality seeking within the own mail, or seeking information in an archive composed by numerous documents in different formats (doc, pdf, etc), or even more as a visual search engine on a set of portals of our interest on simple condition to have them preventively indexed with any type of indexer.

The possible applications are manifold thanks to the great flexibility of this solution. Furthermore, the invasiveness of the system is really minimal and it ensures a good yield.

You can read more about all of these topics by logging into our portal <http://www.vis4you.com/>, where information, contacts and publications relating to the Vis4 World and its products are available.