Dewey Decimal Classification Based Concept Visualization for Information Retrieval

Jae-wook Ahn, Xia Lin, and Michael Khoo

College of Computing and Informatics
Drexel University, Philadelphia, PA USA
{jaewook.ahn,linx,mjk326}@drexel.edu

Abstract. Visual knowledge maps utilizing concepts have great potential to support interactive information retrieval. Unlike keyword-based visual information retrieval, concept-based knowledge maps can make the visualization easier to comprehend and manipulate. In this paper, we introduce our novel visual search interface based on Dewey Decimal Classification concept annotations. The web browser based interface visualizes search results initialized from user queries. Main functions of the interface include interactive manipulation, exploration, and filtering of concepts and links in different levels from overview to details. The visualization connects related concepts not apparent in conventional tree-like hierarchical representations and it can promote discovery of novel concepts during the visual exploration of search space. A real use-case scenario is presented to highlight the advantages of the approach.

Keywords: Concept Visualization, Knowledge Map, Dewey Decimal Classification

1 Introduction

Dewey Decimal Classification (DDC) is a popular document classification system. It has been extensively employed by many traditional libraries to provide users with effective way to browse and search for library resources. It supports a multi-level hierarchy of concept classes that can be used to express the associations and hierarchical relationships of the entire set of resources within any given library. The resource classified with DDC is given a number composed of three or more digits (class, division, and section) that describe the nature of the resource from broader to narrower categories. It is perceived as an efficient way to organize not only in the traditional library settings but also in the modern networked settings [8].

On the other hand, knowledge structure visualization has been actively studied for creating knowledge maps to support interactive search of internet resources. Various approaches have been studied in the literature to present the actual knowledge structure as precise as possible [7]: (1) visualize existing knowledge structures such as tree-like hierarchies (e.g., TreeMap [4]) or ontologies (e.g.,

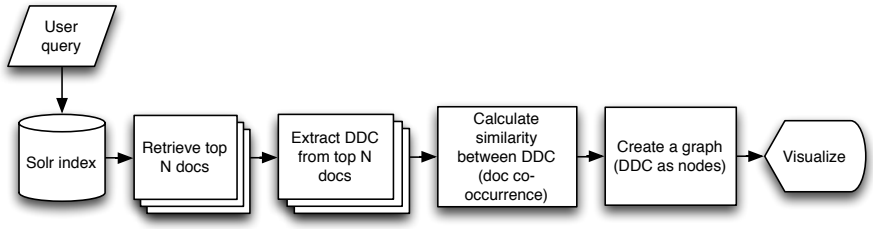

Fig. 1. DDC-based search visualization process

OntoViz [10]), (2) visualize knowledge structures that need to be extracted and learned using various text mining techniques such as automatic thesaurus construction [2] or clustering multi-word terms [9], and (3) visualize knowledge structures through visual metaphors (e.g., [6]).

This paper introduces our novel method to exploit DDC for visualizing knowledge structure of search results. It makes use of existing knowledge structure dynamically derived from live search against three digital library resources. We define four design goals for the knowledge structure visualization: (1) visualize the overview of topics in the search results, (2) visualize concept groups or clusters within the visualization, (3) use DDC to represent concepts, and (4) support discovery of new knowledge using the DDC and visualization. In the following, we will discuss how these goals are achieved from a real use-case of the implemented visualization system.

2 Visualizing Concepts using DDC

The visualization task of this study is based on the Dewey Decimal Classification[1] classes assigned to a large number of digital library records. 263,550 records were collected from three digital libraries (Internet Public Library, Intute, and NSDL)[2] and weighted keywords form their title, description, and subject metadata were used. The resources were matched to multiple DDC classes by calculating the similarity between the resource keyword vectors and the DDC description keyword vectors [5]. Therefore the 263,550 digital library resources were assigned with multiple DDC classes that represent the relevant concepts of the content in three levels – class, division, and section. For example, a web site stored in one of the participating digital libraries is titled as "Olympic History" and presents information about the history of Olympic games. From the automatic DDC class assignment system [5], it was given 10 DDC classes: 796, 943, 945, 942, 949, 941, 940, 948, 944, and 947. The first digits (classes) 7 and 9 represent "Arts & recreation" and "History & geography" respectively. It is clear that

[1] `http://dewey.info`

[2] `http://www.ipl.org`, `http://intute.ac.uk`, `http://nsdl.org`

the resource is understood by the system that it is about sports (i.e., recreation) and history. The remaining two digits (divisions and sections) specifies the detailed topics. For example, the DDC class 796 is labeled as "Athletic & outdoor sports & games," which combines three concepts hierarchically: 7 (class: Arts & recreation), 9 (division: History & geograph), and 6 (section: Technology).

2.1 Visualization Process

By using the DDC classes assigned to the 263,550 records, we implemented a search system that visualizes knowledge structure and concept relationships within search results. Figure 1 depicts the visualization process. We indexed all the documents using Apache Solr information retrieval system[3] so that users can instantly retrieve documents that match their queries. From the retrieved documents, top N documents are selected to calculate the DDC relationships related to them. Each document was assigned with 10 DDC classes following the procedure describe above. Therefore maximum N * 10 (less than N * 10 due to duplicates) DDC classes are retrieved from the database. Then the similarity values between all the DDC class pairs are calculated. Jaccard coefficient is used to calculate the similarity between the DDC class pairs by counting the number of documents that the DDC classes are assigned to.

$$Sim_{Jaccard}(Class_A, Class_B) = \frac{|Class_A Doc \cap Class_B Doc|}{|Class_A Doc \cup Class_B Doc|} \quad (1)$$

Finally a graph is constructed by connecting the DDC classes (nodes) that have higher similarity than a threshold. Similar DDC classes are connected within the graph (links). The resulting graph is visualized in the live search interface using a JavaScript-based network visualization library called Sigma js[4]. ForceAtlas2 [3] force-directed placement graph layout algorithm [1] is used to calculate node locations and the shape of the entire graph.

Figure 2 shows an overview visualization of 694 DDC classes extracted from all documents in the collection. In Figure 3 a user enters a query "olympic AND history" to search for documents that contain both keywords. Among the search results 168 'unique' DDC classes associated with them are visualized. The nodes (discs) represents the DDC 168 classes and the similar nodes (inter-similarity is above a threshold) are linked by curves to each other. The node color reflects the DDC class (first digit) and the legend is located on the left of the screen. Because of the links, groups of similar nodes are clustered together and it can be easily seen that same class nodes (same colors) are forming clusters. For example, there is a large cluster at the top of the screen and is about "social science (DDC=300)." All the search, DDC class retrieval, graph calculation, and visualization functions are performed on the fly when a query is entered.

[3] `http://lucene.apache.org/solr/`
[4] http://sigmajs.org

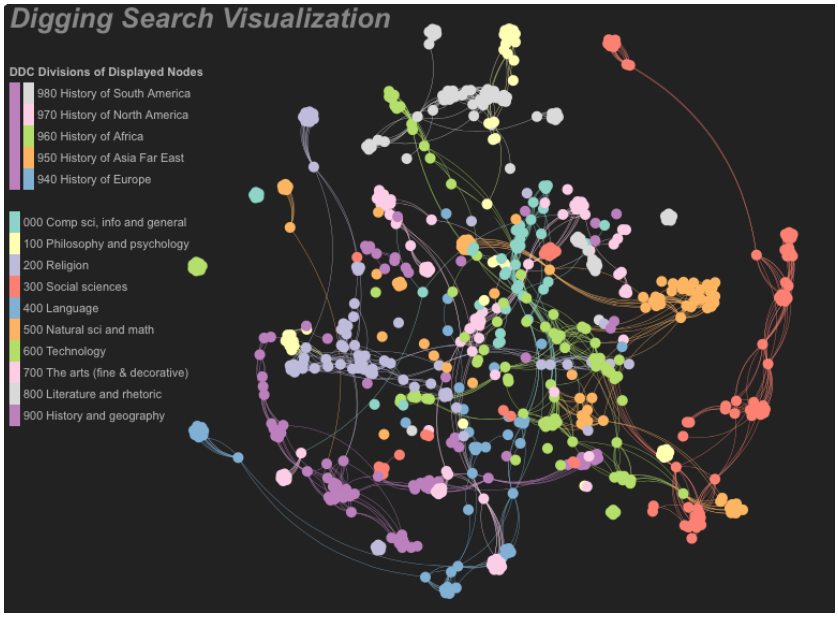

Fig. 2. Overview visualization showing DDC 'class' level distribution of concepts for all 263,550 documents in the collection. 694 classes (nodes) are extracted from the documents and linked to each other based on similarity.

2.2 Additional Visualization Features for Concept Exploration

There are visual features in addition to the basic concept visualization feature to enable more efficient DDC-based search and exploration.

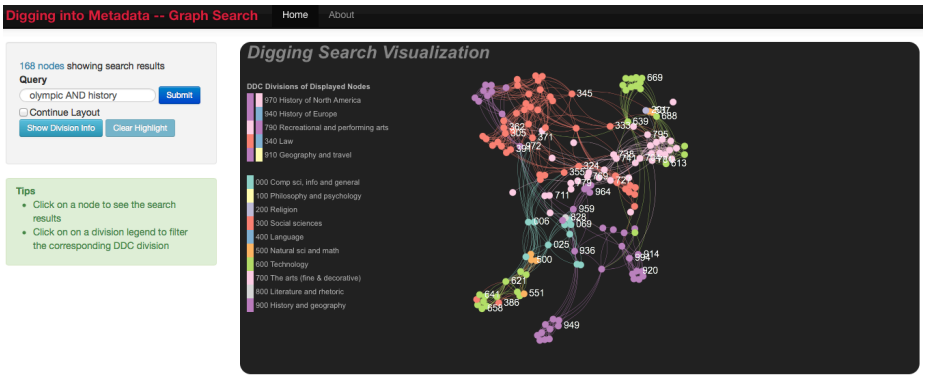

Fig. 3. Overview visualization showing DDC 'class' level distribution of concepts for the query 'olympic AND history'. Overall, history (900-purple), social science (300-red), technology (600-green), and arts (700-light purple) are prevailing. From the division level adaptive legend (top-left) history of two continents (North America and Europe), and 'Recreational Arts' topics are recommended.

Interactive Dynamic Visualization. By using a mouse, users can pan the graph across the screen and zoom in or out to see the details or the overview of the graph. If one clicks on a node a list of documents associated to the concept is displayed (Figure 5).

Adaptive Concept Legend: DDC class-division level. Above the static DDC class legend that show 10 DDC classes and associated colors there is an adaptive DDC class-division legend that is automatically updated following user's panning and zooming action. It counts the most frequent DDC class-division pairs (e.g., DDC=94* with color code purple-pink) of the nodes currently *displayed* on the screen. As the distribution of the DDC class-division pairs can change following the panning and zooming action it shows dynamic information about the area that the user is currently examining (Figure 5).

Filtering classes from legend. If a user clicks on a DDC class and division from the legend, the system highlights the nodes that has the class and the division in red color. It helps users to search for specific group of concepts in a larger concept map and enables targeted examination of concepts (Figure 4).

2.3 DDC-based Visual Search Scenario – Search for "Olympic History"

The design principles of this system is to promote search and browsing by incorporating DDC concepts within a visual search environment. More specifically the following design goals are illustrated in an example search in this section.

1. Visually show the overview of concepts included in a search result.
2. Help users search for specific concepts or *concept groups* easily within the visualization.
3. Help users find relevant documents by examining the DDC concepts or *concept groups* mapped in the visualization.
4. Support users discover new information by following links between *concept groups*. The *concept groups* are formed by linking homogeneous concepts but heterogeneous groups are inter-linked as well.

We will show how these goals are achieved in a real use-case information retrieval scenario. Suppose a user was assigned a task to find out as many DDC classes and documents as possible to write a report about "Olympic history." Using the search interface, she enters a query "olympic AND history." The query is immediately transmitted to the backend Solr index and retrieves 103 documents from the entire collection (236,550 documents). From the 103 documents 168 DDC classes are identified by looking up the database that stores the pre-calculated DDC classes for all documents in the dataset. The system then calculates the Jaccard similarity values of 5,253 DDC class pairs (168 * 167 / 2) by calculating the co-occurrence of the documents they are assigned to. The DDC

classes with higher similarity values above a specific threshold are linked and creates a graph (Figure 3).

As described before the nodes are color-coded by their DDC classes and the clusters are easily identified from the map (Goal 1 and 2). By clicking on the adaptive legend (top-left of the screen) the selected DDC class-division (first and second digits) classes are highlighted in the visualization. In Figure 4, the adaptive legend shows that the most frequent DDC class-division pairs are 970:History of North America, 940:History of Europe, and so forth. A user clicks on DDC 940 then the nodes starting with 94 are all highlighted in red and the other nodes are de-highlighted. A node directly connected to the highlighted cluster is not de-highlighted and retains its original color (purple, DDC=936), which is a potentially relevant DDC class worth further examination (Goal 3).

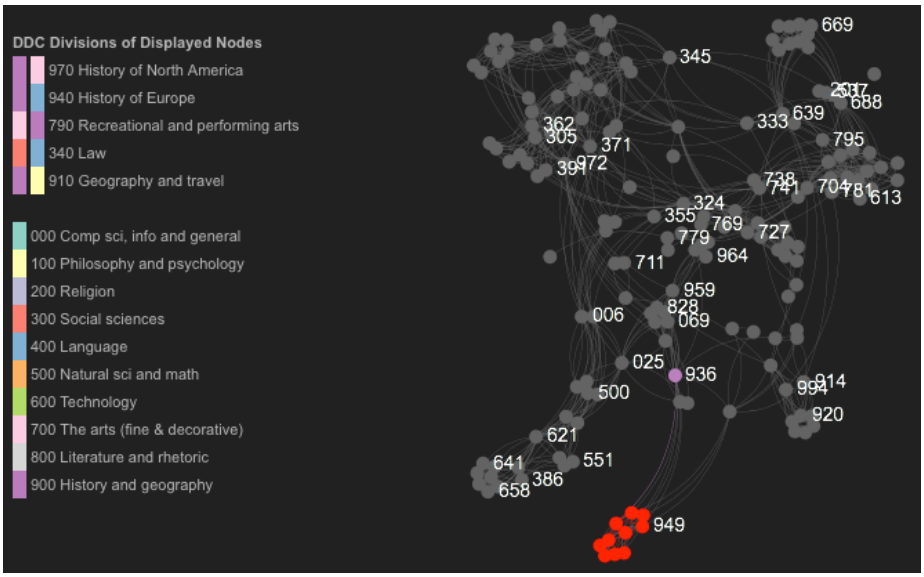

Fig. 4. From the legend, any topic can be selected and highlighted within the map. In this example 940 – History of Europe is selected and the corresponding nodes are highlighted (in red). Remaining nodes are dimmed. Using this feature users can easily select concepts and discover how they are connected to each other.

The visualization can recommend novel concepts that are difficult to discover using conventional search and browsing methods. Figure 5 shows the new discovery of concepts that are relatively further in the DDC hierarchy. It zooms into the DDC=790: Recreation and performing arts area in the map, which is intuitive enough to anticipate that sports related information should be under this category. It shows 10 DDC=790 classes (DDC=790, 791, 792, 793, etc. in red color) that the user might have intended to examine in the first place. In addition

to them there are several connected classes that are not under the 790 category. The system can lead users to examine DDC=600 concepts (i.e., 613, 617, and 636 in green) which are under the broader technology classes (DDC=600) but connected to the initial 790 classes. By examining the documents under one of the 600 concepts (Figure 5) it can be verified that one of the sites "SR: Olympic Sports" contains relevant information about Olympic games statistics and history. Because DDC=613 is under the technology category (DDC=600) that may be seemingly unrelated to the search task, it may be challenging for a user to be motivated to examine the category. Only after starting the visual exploration of the chains of DDC classes from easier ones (i.e., 700: Recreation), more difficult and novel categories can be discovered (Goal 4).

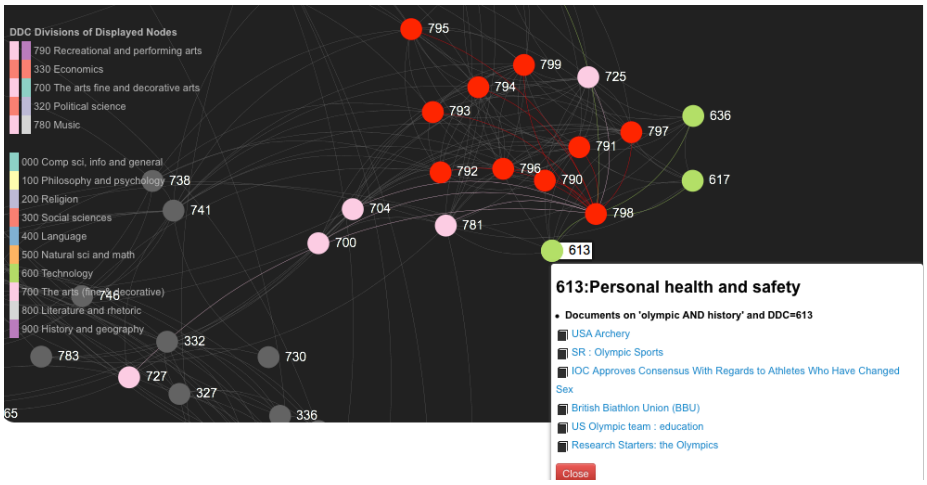

Fig. 5. Zoomed into '790: Recreation and performing arts' and their neighbor area (top-right from the overview visualization). The legend is updated accordingly and shows new clusters of concepts. There are three Technology (600) related nodes (green) and the documents about the concepts can be examined by clicking on the node. It discovers a new document "SR: Olympic Sports" that may be relevant to the task. By clicking on the document title one can open the website and can confirm it contains Olympic related statistics and history.

3 Conclusions

In this paper we introduced a visual information retrieval approach based on DDC classification. We annotated 263,550 records from three digital libraries with automatically generated DDC classes and implemented a information retrieval system featuring a graph visualization that connects similar DDC concepts based on the document co-occurrence between them. We showed the advantages of our approach by a real use-case scenario. The example demonstrated

that the approach could support interactive and dynamic knowledge visualization and could promote the discovery of concept clusters and unknown concepts. Our future research plans include a full-fledged user study to learn about the advantages and disadvantages of the approach from real users.

References

1. Eades, P.: A heuristic for graph drawing. Congressus Numerantium 42, 149–160 (1984)
2. Grefenstette, G.: Explorations in automatic thesaurus discovery. Springer (1994)
3. Jacomy, M., Heymann, S., Venturini, T., Bastian, M.: Forceatlas2, a continuous graph layout algorithm for handy network visualization. Medialab center of research 560 (2011)
4. Johnson, B., Shneiderman, B.: Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In: VIS '91: Proceedings of the 2nd conference on Visualization '91. pp. 284–291. IEEE Computer Society Press, Los Alamitos, CA, USA (1991), http://portal.acm.org/citation.cfm?id=949654
5. Khoo, M., Tudhope, D., Binding, C., Abels, E., Lin, X., Massam, D.: Towards Digital Repository Interoperability: The Document Indexing and Semantic Tagging Interface for Libraries (DISTIL). In: Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F. (eds.) Theory and Practice of Digital Libraries, Lecture Notes in Computer Science, vol. 7489, pp. 439–444. Springer Berlin Heidelberg (2012), http://dx.doi.org/10.1007/978-3-642-33290-6_49
6. Kleiberg, E., Van De Wetering, H., Van Wijk, J.J.: Botanical visualization of huge hierarchies. In: Information Visualization, IEEE Symposium on. pp. 87–87. IEEE Computer Society (2001)
7. Lin, X., wook Ahn, J.: Challenges of knowledge structure visualization. In: Internatinal UDC Seminar 2013 (2013)
8. Saeed, H., Chaudhry, A.S.: Using dewey decimal classification scheme (ddc) for building taxonomies for knowledge organisation. Journal of Documentation 58(5), 575–583 (2002)
9. SanJuan, E., Ibekwe-SanJuan, F.: Text mining without document context. Information Processing & Management 42(6), 1532–1552 (2006)
10. Sintek, M.: Ontoviz, http://protegewiki.stanford.edu/wiki/OntoViz