Wikipedia Category Structure in the Light of Dewey Decimal Classification System: an Evaluation

By
Piyali Ghosh

Thesis submitted for the award of the Faculty of Arts of Jadavpur University in partial fulfilment of the requirements for the degree of **Doctor of Philosophy** in Library & Information Science

Department of Library and Information Science
Jadavpur University
Kolkata – 700032
2024

Chapter 1: Introduction

Wikipedia becomes adult member of information resources which provides multilingual semi-structured content. Wikipedia is being used to classify web content, is used for creating large scale taxonomies, it is very natural to find answer of a question how Wikipedia itself organise its content or what is the technology or process used for.

1.1. The aim of this study

The proposed study will evaluate this category structure of Wikipedia, to explore the classification system and the arrangement patterns of pages exits in this vast mass generated encyclopedia and to visualize the category structure exists in this encyclopedia.

Chapter 2: Literature Review

The collected articles were filtered by three main themes, like ‘Works on evaluation of Wikipedia as a knowledge organization tool’, ‘Wikipedia category structure’, and ‘comparison between Wikipedia and classification tools’ (see below figure).

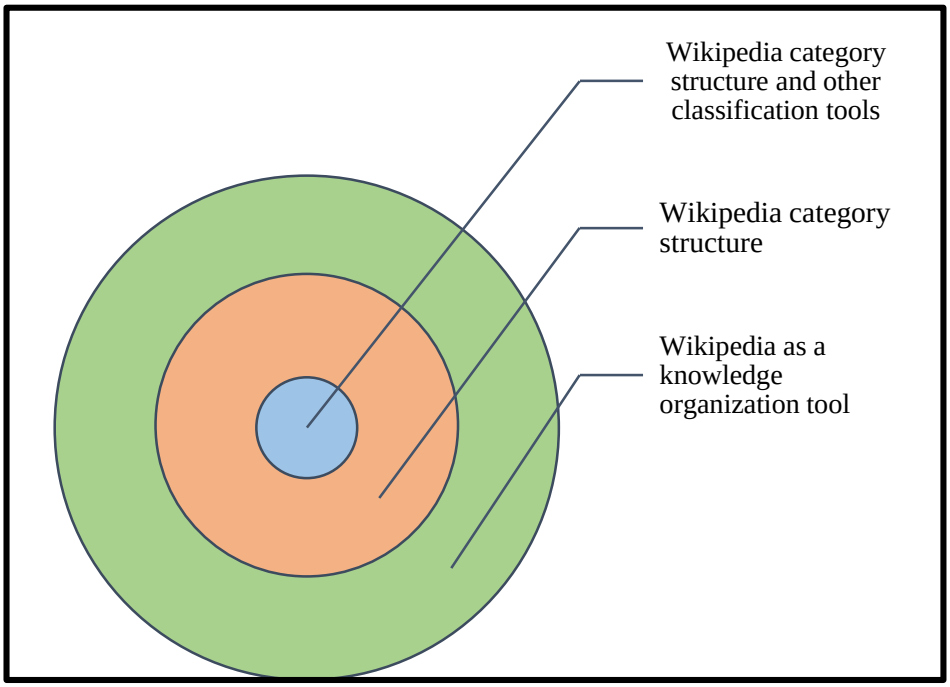

Figure 1: Thematic distributions of collected literature

2.1. Literature on "Wikipedia as a knowledge organization tool"

According to Voss (2009) Wikipedia is being or may be used as knowledge organization tool. Wikipedia's category network may be used as a tool for document indexing (Chahine et al., 2011).

2.2. Literature on "Wikipedia category structure"

An evolution of Wikipedia category structure by Suchecki, Salah and Gao (2012) focused on the clustering of articles exist in Wikipedia defined by its category system. To know about inter-relation between categories a research work (Szymanskki & Duch, 2012) designed visualization map.

2.3. Literature on "Wikipedia category structure and other classification tools"

Now Wikipedia is being used for categorising web videos (Chen et al., 2010), different web resources (Medeiros et al., 2018), learning resources (Marek, Rensing, & Steinmetz, 2018). Wikipedia can also be utilised for improving document classification (Wang et al., 2009) for classifying short text of Tweet or any micro blogging sites.

2.3. Observations

Wikipedia is now using to classify other web contents. Therefore Wikipedia category structure should be examined with any existing most popular classification scheme. The above literature review shows a lack in this regard. More research work should be done on this issue where Wikipedia's category structure could be scrutinized on the basis of any classification scheme.

Chapter 3: Research Design

3.1. Statement of the problem

The problem of the proposed research is: **The evaluation of Wikipedia's category structure**

3.2. Significance of the problem

Evaluation is needed to know and to explore the classification system and the arrangement patterns of pages exits in this vast mass generated encyclopedia. Evaluation is also needed to visualize the category structure exists in this encyclopedia.

3.3. Research questions

The following research questions are designed:

1. Is there any similarity between the categories structure found in Wikipedia with the structure of DDC?
2. Is the classification of Wikipedia categories into one or more categories influenced by DDC classification system?

3.4. The Whole Research Design

The next table shows whole research design including research questions, variables, methodologies and units of measurements in DDC and in Wikipedia.

Table 1: Research Design

SN	Research Questions	Variables	Methodology	Units of measurements in DDC	Units of measurements in Wikipedia
1.	Is there any similarity between the categories structure found in Wikipedia with the structure of DDC?	Resemblance	One to one mapping of the section names of the each divisions of the main classes "philosophy", "social sciences", "science" of DDC with Wikipedia categories	Section names of the each divisions of the main classes "philosophy", "social sciences", "science"	Similar name of the DDC sections of the each divisions of the main classes "philosophy", "social sciences", "science"
2.	Is the classification of Wikipedia categories into one or more categories influenced by DDC classification system?	Influence	Comparing the parent categories of the Wikipedia category with DDC	DDC divisions of the main classes "philosophy", "social sciences", "science"	Wikipedia categories with the same name of DDC divisions of the main classes "Philosophy", "Social Sciences", "Science"

Chapter 4: Data Analyses

4.1. Data analyses on 'Resemblance'

To analyse the resemblance or similarity, the sections under the main classes "Philosophy", "Social Sciences", "Science" mentioned in DDC 23rd edition were noted down. The following table represents the assigned sections of DDC and the availability of categories of same name in Wikipedia.

Table 2: Resemblance between two knowledge organization tools DDC and Wikipedia

	Sections assigned in DDC	Fully similar categories in Wikipedia	Partially Similar categories in Wikipedia	Dissimilar categories in Wikipedia
Philosophy	89	58 (65%)	11 (12%)	20 (23%)
Social Sciences	90	46 (51%)	13 (15%)	31 (34%)
Science	93	47 (50%)	08 (9%)	38 (41%)
	272	151 (55%)	32 (12%)	89 (33%)

The above data and the below pie charts show that Wikipedia contains maximum similar categories for the subject "Philosophy", and minimum similar categories for the subject "Science".

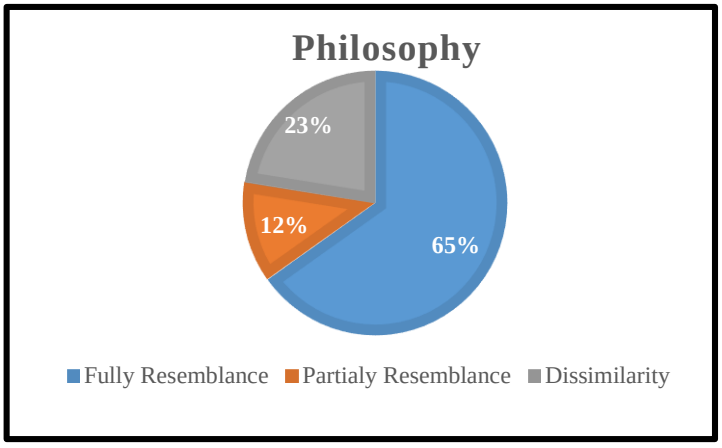

Figure 2: Resemblance in DDC main class Philosophy

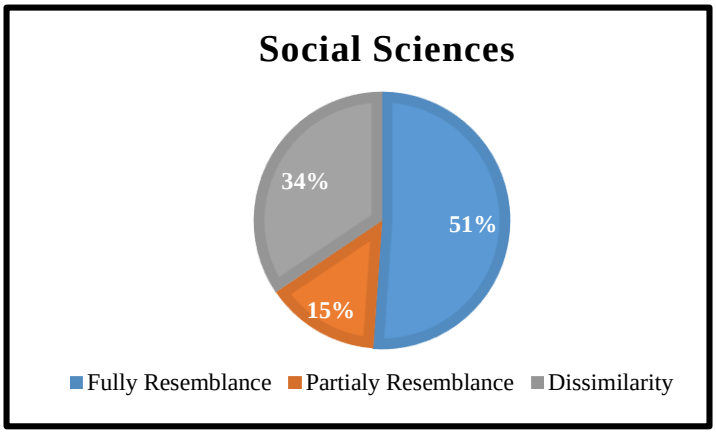

Figure 3: Resemblance in DDC main class Social Sciences

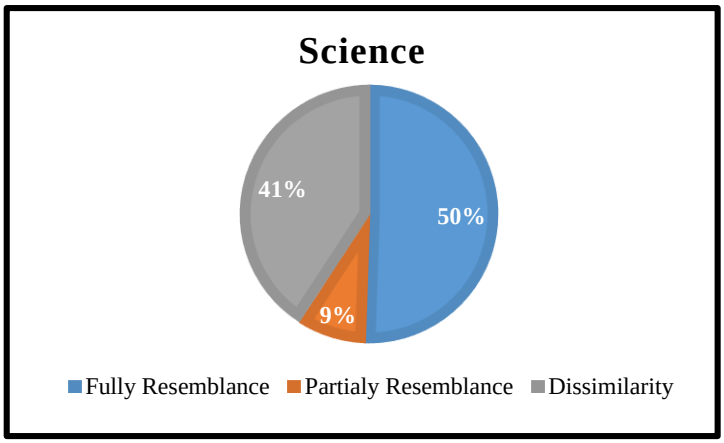

Figure 4: Resemblance in DDC main class Science

Total 272 sections are noted from DDC 23rd. Out of 272 sections, **151 sections (55%) were found in Wikipedia as category. 89 sections (33%) were not** found in Wikipedia. There is partial similarity of categories for 12%, it means 32 categories were found partially similar to the collected sections. See the below pie chart:

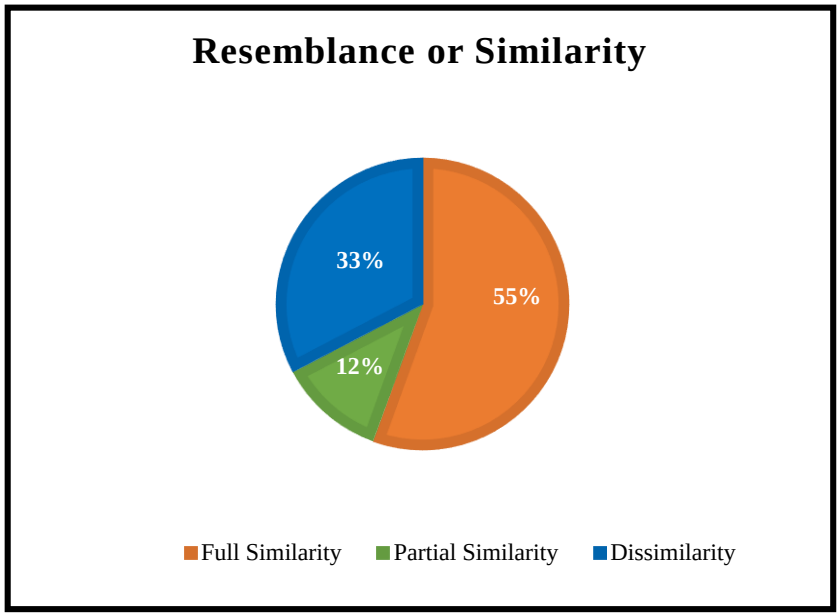

Figure 5: Resemblance between Wikipedia categories and DDC

4.2. Influence of DDC Classification Scheme on Wikipedia

It was decided to find out whether there is any influence of DDC on Wikipedia or not by comparing the parent categories of the Wikipedia category with DDC.

Table 3: Marking Scheme for the data analyses on "Influence"

Marking Schemes	**Marks**
When the category is directly under the category Philosophy/Social Science/Science	**1**
When the category is one category away from the parent category similar to Philosophy/Social Science/Science	**2**
When the category is two categories away from the parent category similar to Philosophy/Social Science/Science	**3**
When the category is three categories away from the parent category similar to Philosophy/Social Science/Science	**4**
When Parent category is other than Philosophy/ Social Science/Science	**0**

On the basis of above marking scheme, points was given to the sample categories similar to Philosophy/Social Sciences/ Science divisions. It was to find out is there any influence of DDC classification scheme on Wikipedia category structure or not. The above marking scheme says if there is any influence then points would be "1" or near to 1. If the point is "0" then it means there is no influence of DDC scheme on Wikipedia category structure. The results are as follows:

4.2.1. Philosophy Divisions

Table 4: Points gained by the Categories similar to Philosophy Divisions

Categories similar to Philosophy Divisions	**Points gained**
Metaphysics	2
Epistemology	2
Parapsychology	0

Philosophical schools	1
Psychology	0
Philosophical logic	4
Ethics	2
Eastern Philosophy	4
Total categories 8 similar to philosophy divisions	Total points15

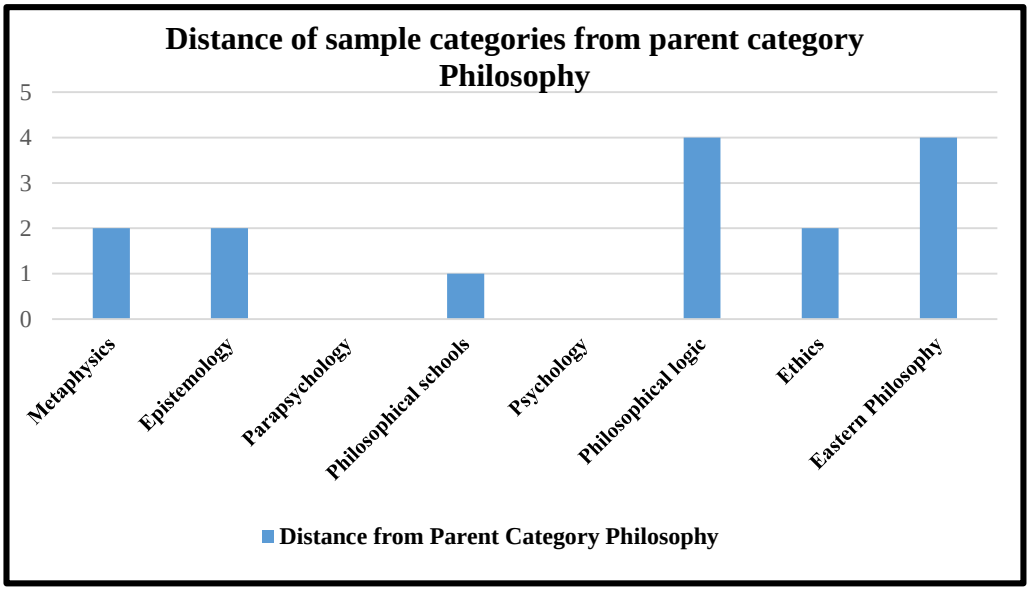

Figure 6: Distance of sample categories from parent category Philosophy

4.2.2. Social Sciences Divisions

Table 5: Points gained by the Categories similar to Social Sciences Divisions

Categories similar to Social Science Divisions	Points gained
Statistics	0
Political Science	1
Economics	1
Law	4
Public administration and military science	1

Education	0
Commerce, communications, transportation	0
Customs, etiquette, folklore	2
Total categories 8 similar to philosophy divisions	Total points 09

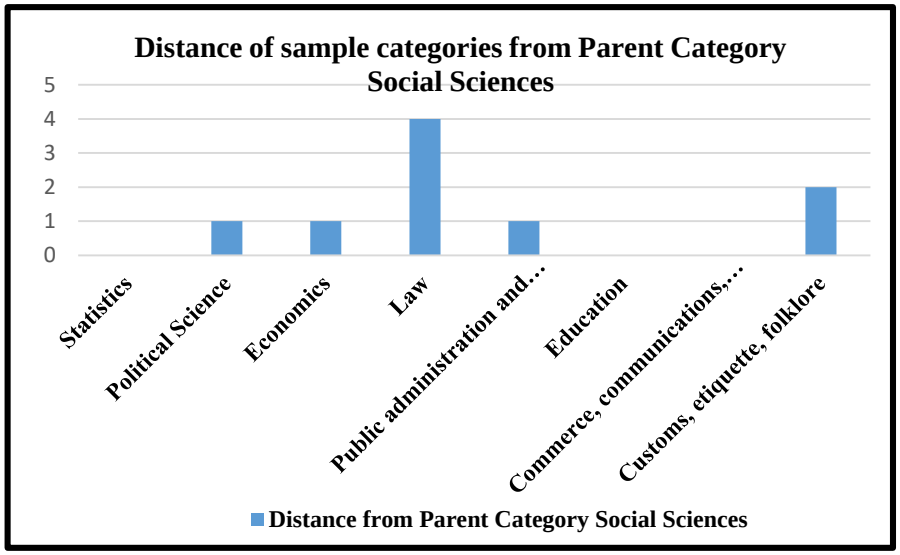

Figure 7: Distance of sample categories from parent category Social Sciences

4.2.3. Science Divisions

Table 6: Points gained by the Categories similar to Social Sciences Divisions

Categories similar to Social Science Divisions	Points gained
Statistics	0
Political Science	1
Economics	1
Law	4
Public administration and military science	1
Education	0

Commerce, communications, transportation	0
Customs, etiquette, folklore	2
Total categories 8 similar to philosophy divisions	Total points 09

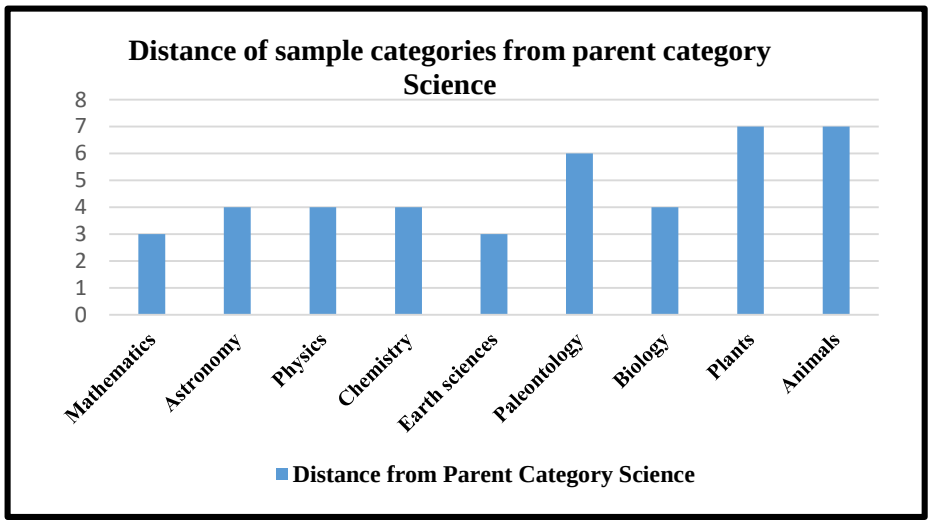

Figure 8: Distance of sample categories from parent category Science

Chapter 5: Findings

5.1. Findings on the basis of the variable "Resemblance"

1) **55% Similar Categories:** For three main classes namely "Philosophy", "Social Sciences", "Science", Wikipedia contains 55% **similar categories.** Out of selected 272 sections of DDC, 151 sections were found in Wikipedia as category.
2) **33%Dissimillar Categories:** Out of selected 272 sections of DDC, 89 sections were not found in Wikipedia as category. The percentage of dissimilarity is 33%.
3) **12% Partially Similar Categories:** Few categories were there which are not fully but partially similar with the DDC sections.
4) **Maximum and Minimum Similar categories:** Wikipedia contains maximum similar categories for the subject "Philosophy" which is 65%, and minimum similar categories for the subject "Science" which is 50%. The encyclopedia contains 51% similar categories for the subject "Social Sciences".

5.2. Findings on the basis of the variable "Influence"

1) To find out the result a marking scheme was designed, which indicates that if the category hierarchy of a particular subject is influenced by DDC classification scheme then it must have smallest point, with no zero.
2) Two zeroes were found in the sample categories collected similar to DDC philosophy divisions. The total point is 15.
3) Three zeroes were found in the sample categories collected similar to DDC social sciences divisions. Here the total point is 09.
4) Apparently the category structure of the subject "Science" is found not much influenced by DDC classification scheme, as the total point here is 42. But all the sample categories collected similar to DDC sciences divisions are having parent category "Science". There is no zero. The lowest point is "3" gained by "Mathematics" and "Earth Science" and the highest point is "7", gained by "Plants" and "Animals".
5) Out of 25 sample Wikipedia categories (similar to DDC Philosophy, Social Sciences, Science divisions) 5 categories are not having the same parent categories as described in DDC. Those 5 categories (with 0 points) are Parapsychology, Psychology, Statistics, education, Commerce, communications, transportation. Rest of the 20 categories are having similar parent categories as described in DDC (see next pie chart).

5.3. Findings of Research Questions

The following table accumulate the findings of the research questions designed for the study.

Table 7: Research questions and findings

SN	Research Questions	Methodology	Findings
1.	Is there any similarity or resemblance between the categories structure found in Wikipedia with the structure of DDC?	One to one mapping of the subclasses/division under main classes "philosophy", "social sciences", "science" of DDC with Wikipedia	Yes, there is a similarity between Wikipedia category structure and DDC class structure. Wikipedia contains **55% similar categories.** Out of selected **272** sections of DDC, **151** sections were found in Wikipedia as category.
2.	Is the classification of Wikipedia categories into one or more categories influenced by DDC classification system?	Comparing the parent categories of the Wikipedia category with DDC	Wikipedia is somehow influenced by DDC classification system. Out of 25 DDC divisions, 5 divisions are not found in Wikipedia with the

			same parent category as DDC. **80 % of total sample categories are influenced by DDC class divisions**, as they are having similar parent categories described as in DDC.

Chapter 6: Conclusions

The analyses on the basis of "Resemblance" variable give a result where the sample categories of three subjects are resemblanced with DDC class names. The percentage of resemblance is 55. Out of selected 272 sections of DDC, 151 sections were found in Wikipedia as category.

The analyses on the basis of variable "Influence" shows that the sample categories are having parent categories similar to DDC class structure. The hierarchy of Wikipedia categories sub-categories is influenced by DDC. Out of 25 sample Wikipedia categories (similar to DDC Philosophy, Social Sciences, Science divisions) 5 categories are not having the same parent categories as described in DDC. Here the influence of DDC class structure on Wikipedia category structure is 80%.

The present study founds Wikipedia's similarity with DDC classification scheme and in many places the influence of DDC class structure is found in Wikipedia. Therefore it may be concluded that somehow accidently or incidentally Wikipedia category structure is similar with DDC class structure and there is an influence of DDC in Wikipedia. Further research is required to know the influence of other scheme on this mass generated encyclopedia.

6.1. Scope for Further Research

During the research works, the present study finds following thesaurus like elements in Wikipedia:

Table 8: Thesaurus like features in Wikipedia

Elements	**Thesaurus like features**
Redirects	"see" references by linking synonyms to preferred terms.
Disambiguation pages	Homonyms.
The poly hierarchical category structure	Broader and narrower term relationships
Links between pages	Related term indicators

Because of these thesaurus like features many researchers uses Wikipedia for metadata enrichment, text clustering, and classification. When this encyclopedia is being used to classify documents, resources, then it is necessary to search that is there any influence of any classification scheme on Wikipedia or not. The present study found Wikipedia's similarity with DDC classification scheme. But the research work may be done for other classification scheme also.

6.2. Recommendation

There are few recommendations:

- The entire collection of the encyclopedia may be classified according to any popular classification scheme.
- There is many loop connections between all the connected categories. The collection of categories may be organised in tree like hierarchy.
- As this encyclopedia is the biggest encyclopedia of the world, anyone can lost in its collection. Therefore it is recommended to design proper site map of categories (without hidden loops) to guide its users all over the world.

N.B. The original thesis contains the detailed bibliography and it has not been included in this synopsis and in the reference portion below only the references of the literature review part of this synopsis have been included.

References

Chahine, C. A., Chaignaud, N., Kotowicz, P., & Pécuchet, J. (2011). Conceptual indexing of documents using Wikipedia. *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (pp. 195-202). IEEE Xplore Digital Library. https://ieeexplore.ieee.org/document/6040518

Chen, Z., Cao, J., Song, Y., Zhang, Y., & Li, J. (2010). Web video categorization based on Wikipedia categories and content-duplicated open resources. *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference* (pp.1107-1110). ACM Digital Library. https://doi.org/10.1145/1873951.1874162

Marek, M., Rensing, C. and Steinmetz, R. (2007). *Categorizing learning objects based on Wikipedia as substitute corpus.* https://www.researchgate.net/publication/221549919_Categorizing_Learning_Objects_Based_On_Wikipedia_as_Substitute_Corpus

Medeiros, J. F., Nunes, B. P., Siqueira, S. W. M., & Leme, L. A. P. (2018). *Tag the web: using Wikipedia categories to automatically categorize resources on the web.* https://2018.eswc-conferences.org/files/posters-demos/paper_275.pdf

Suchecki, K., Salah, A., & Gao, C. (2012). Evolution of Wikipedia's category structure. *WSPC/Instruction File.* https://arxiv.org/pdf/1203.0788.pdf.

Szymanski, J., & Duch, W. (2012). Self organizing maps for visualization of categories. *In Neural Information Processing, ICONIP 2012, Lecture Notes in Computer Science.*7663. https://link.springer.com/chapter/10.1007/978-3-642-34475-6_20.

Voss, J. (2009). Wikipedia as knowledge organization system. *International Cataloguing and Bibliographic Control: Quarterly Bulletin of the IFLA UBCIM Programme, 39*(2), 41-42. https://search.proquest.com/openview/7deb2f5435b182877a1ee7557816261d/1?pq-origsite=gscholar&cbl=60376

Wang, P., Hu, J., Zeng, H., & Chen, Z. (2009). Using Wikipedia knowledge to improve text classification. *Knowledge and Information Systems, 19*, 265-281. https://www.semanticscholar.org/paper/Using-Wikipedia-knowledge-to-improve-text-WanHu/fc1d23d2f9167d13ef1bce098ef55d1b40894dd4