A Knowledge-graph based Taxonomy Construction Method

András London[1,2], János Zsibrita[2] and Rio Fear[3]

University of Szeged, Institute of Informatics, Arpád tér 2, H-6720, Szeged Hungary

Black Swan Data, Zászló utca 4, H-6722 Szeged, Hungary

Black Swan Data, 15th Floor, 10 York Rd, London SE17ND, United Kingdom

Abstract

A **taxonomy** is a hierarchically organized categorization of concepts or entities, for example a Wikipedia category, an ACM Classification System, or an Amazon Product Category. For a great many companies around the world domain-specific taxonomies form a crucial component of the provision data-driven solutions: they can help in search optimization, browsing, organization and storage of information, and much more besides. Automated taxonomy building has been well researched in the recent years. Most approaches apply NLP tools to a text corpus e.g. [1], some of them utilize knowledge-graphs, e.g. [2], like Wikipedia or WordNet, while others combine the previous approaches e.g. [3].

In this work, we

- provide a simple, Wikipedia knowledge graph-based methodology;
- regard the taxonomy construction as a series of basic graph algorithms;
- on case-studies we present the performance of the procedure.

Wikipdia graph

We construct the **Wikipedia-based knowledge graph** proposed and deployed by Aspert et al. [4] available at `https://lts2.epfl.ch/Datasets/Wikipedia/`. This graph is a **directed multigraph** with multiple nodes and edge types:

- **Node** $\longleftrightarrow$ a Wikipedia article or a Wikipedia category
- **Links_to** type edge $\longleftrightarrow$ a hyperlink between two articles
- **Belongs_to** type edge $\longleftrightarrow$ hyperlink between an article or (sub)category node and a category

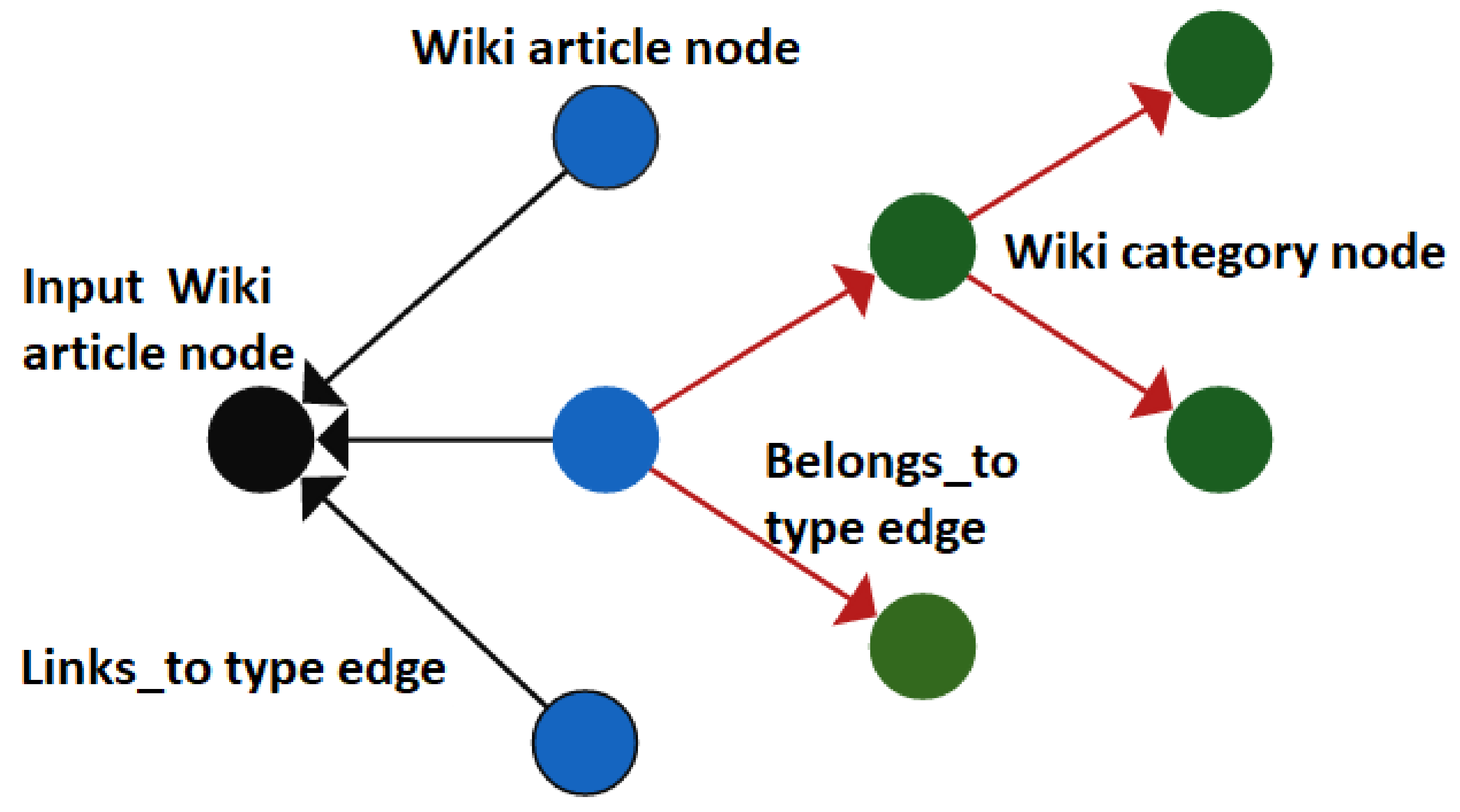

Figure 1: Wikipedia graph structure. Blue (black) nodes: articles (as input). Green nodes: category pages. Black edges: hyperlinks connecting articles. Red edges: hyperlinks connecting articles or subcategories and parent categories.

Taxonomy entity selection

The **taxonomy generator** is initialized with a collection of Wikipedia article type nodes $P = \{P^1, P^2, \dots\}$ and Wikipedia category type nodes $C = \{C^1, C^2, \dots\}$ which we process according to the following steps:

1. Construct a set $\mathcal{P} = \{P_1^1, P_2^1, \dots; P_1^2, P_2^2, \dots; \dots\}$ of all nodes which have a 'link_to' edge to one or more of the input pages, P.
2. Start a depth-first traversal over each node $P_i^j \in \mathcal{P}$ for all 'belongs_to' type outgoing edges from P_i^j. At the first level this will result in the set $C_i^j = \{C_{i,1}^j, C_{i,2}^j, \dots\}$ of categories which the page P_i^j "belongs_to", at the second level the set of higher 'super'-categories of categories in C_i^j will be reached, and so on.)
   2.1 If for a category node $C_{i,k}^j$, found during the traversal process $C_{i,k}^j \in C$ is satisfied, then add P_i^j to a "filtered" entity list $\mathcal{L}$;
   2.2 Else, go to step 2, until all elements of $\mathcal{P}$ have been iterated over.

Some technical notes:

- After step 1. a fast filtering procedure can be applied by simply deleting any nodes from set $\mathcal{P}$ for which the node's corresponding Wikipedia page name either begins with a number (i.e. "2019_in_tennis") or contains the terms "by_year", "of_the_year", "List_of", or "_in_" (i.e. "Tennis_in_Hungary").
- In step 2. a stop criteria is required to restrict the maximum depth of the traversal process due to performance issues. In our experiments the criteria was set to a maximum depth level of four starting from the root node, provided that a category page in C had not already been reached.

Taxonomy creation

The final step is to classify each entity $e \in \mathcal{L}$ with a category and to provide a **hierarchical category organization**.

1. For each e let C^e be the set of categories which e belongs to, that is, the neighborhood of e based on its outgoing 'belongs_to' type edges. Note that C^e is determined in step 2 of the entity extraction process.
2. Let $\mathcal{C}$ be the set of all distinct categories in $\bigcup_{e \in \mathcal{L}} C^e$. We define a bipartite graph over the disjoint node sets $\mathcal{L}$ and $\mathcal{C}$, where $e \in \mathcal{L}$ and $c \in \mathcal{C}$ are connected if e belongs to category $c \in \mathcal{C}$.
3. Starting from c with the highest degree we greedily assign entities to categories step-by-step by removing the assigned entities and corresponding category in each step.
4. Finally, to organize categories into a proper hierarchy one may use a pruning heuristics used e.g. in [5].

A case-study and evaluation

Domain-specific taxonomies are usually evaluated either by comparing them to manually-built (Gold Standard) taxonomies or by requesting feedback from experts in the field. One of our case-studies is targeted to build a taxonomy covering various sports.

- For example, for tennis the input Wikipedia page nodes were *Tennis*, *Association_of_Tennis_Professionals* and *Women's_Tennis_Association*, while the input category page node was *Tennis*.

Table 1 and 2 shows our **experimental results regarding coverage** (recall) values comparing the Gold Standard and automated taxonomy methods.

Taxonomy/category	TP/ Gold Players	TP/ Gold Teams
American football	**90.65** (97/107)	**100** (32/32)
Basketball	**89.15** (403/452)	**100** (30/30)
Motorsport	**88.38** (784/88)	–
Soccer	**79.8** (399/500)	**48.83** (294/602)
Tennis	**75.45** (206/273)	–

Table 1: Coverage (ratio of true positives of automatically extracted entities and manually defined gold standard entities) results for several sports related taxonomies.

Taxonomy/category	TP/Gold All cat.	All Auto
American football	**74.59** (138/185)	4,068
Basketball	**89.67** (443/494)	5,526
Motorsport	**86.12** (807/937)	5,862
Soccer	**62.96** (731/1161)	3,096
Tennis	**67.17** (262/390)	2,077

Table 2: Coverage (ratio of true positives of automatically extracted entities and manually defined gold standard entities) results for several sports related taxonomies.

- The automated method finds many more relevant entities than the Gold Standard.
- However, for the purposes of this investigation this is a secondary concern to the primary aim of achieving a high recall compared to the Gold Standard.
- The high-precision reduction of irrelevant entities from the auto taxonomy (false positives) remains for future work.

References

(1) Hearst, M. A. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, 1992, pp 539–545.

(2) Ponzetto, S. P.; Strube, M. In *AAAI*, 2007; Vol. 7, pp 1440–1445.

(3) Ponzetto, S. P.; Strube, M. *Artificial Intelligence* **2011**, *175*, 1737–1756.

(4) Aspert, N.; Miz, V.; Ricaud, B.; Vandergheynst, P. In *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp 1188–1193.

(5) Medelyan, O.; Manion, S.; Broekstra, J.; Divoli, A.; Huang, A.-L.; Witten, I. H. In *Extended Semantic Web Conference*, 2013, pp 367–381.

inf.u-szeged.hu/∼london

December, 2019

london@inf.u-szeged.hu