A Web-scale system for scientific knowledge exploration

Zhihong Shen
Microsoft Research
Redmond, WA, USA

Hao Ma
Microsoft Research
Redmond, WA, USA

Kuansan Wang
Microsoft Research
Redmond, WA, USA

`{zhihosh,haoma,kuansanw}@microsoft.com`

Abstract

To enable efficient exploration of Web-scale scientific knowledge, it is necessary to organize scientific publications into a hierarchical concept structure. In this work, we present a large-scale system to (1) identify hundreds of thousands of scientific concepts, (2) tag these identified concepts to hundreds of millions of scientific publications by leveraging both text and graph structure, and (3) build a six-level concept hierarchy with a subsumption-based model. The system builds the most comprehensive cross-domain scientific concept ontology published to date, with more than 200 thousand concepts and over one million relationships.

1 Introduction

Scientific literature has grown exponentially over the past centuries, with a two-fold increase every 12 years (Dong et al., 2017), and millions of new publications are added every month. Efficiently identifying relevant research has become an ever increasing challenge due to the unprecedented growth of scientific knowledge. In order to assist researchers to navigate the entirety of scientific information, we present a deployed system that organizes scientific knowledge in a hierarchical manner.

To enable a streamlined and satisfactory semantic exploration experience of scientific knowledge, three criteria must be met:

- a comprehensive coverage on the broad spectrum of academic disciplines and concepts (we call them *concepts* or *fields-of-study*, abbreviated as *FoS*, in this paper);
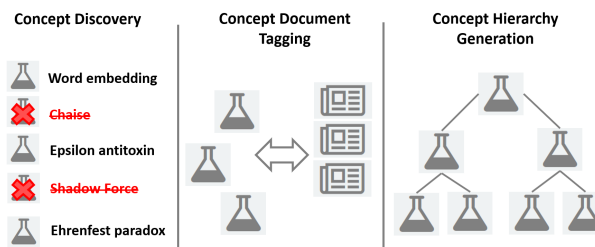

Figure 1: Three modules of the system: *concept discovery*, *concept-document tagging*, and *concept hierarchy generation*.

- a well-organized hierarchical structure of scientific *concepts*;

- an accurate mapping between these *concepts* and all forms of academic *publications*, including books, journal articles, conference papers, pre-prints, etc.

To build such a system on Web-scale, the following challenges need to be tackled:

- **Scalability:** Traditionally, academic discipline and concept taxonomies have been curated manually on a scale of hundreds or thousands, which is insufficient in modeling the richness of academic concepts across all domains. Consequently, the low concept coverage also limits the exploration experience of hundreds of millions of scientific publications.

- **Trustworthy representation:** Traditional concept hierarchy construction approaches extract concepts from unstructured documents, select representative terms to denote a concept, and build the hierarchy on top of them (Sanderson and Croft, 1999; Liu et al., 2012). The concepts extracted this way not only lack authoritative definition, but also

87

	Concept discovery	Concept tagging	Hierarchy building
Main challenges	scalability / trustworthy representation	scalability / coverage	stability / accuracy
Problem formulation	knowledge base type prediction	multi-label text classification	topic hierarchy construction
Solution / model(s)	Wikipedia / KB / graph link analysis	word embedding / text + graph structure	extended subsumption
Data scale	$10^5 - 10^6$	$10^9 - 10^{10}$	$10^6 - 10^7$
Data update frequency	monthly	weekly	monthly

Table 1: System key features at a glance.

contain erroneous topics with subpar quality which is not suitable for a production system.

- **Temporal dynamics:** Academic publications are growing at an unprecedented pace (about 70K more papers per day according to our system) and new concepts are emerging faster than ever. This requires frequent inclusion on latest publications and re-evaluation in tagging and hierarchy-building results.

In this work, we present a Web-scale system with three modules—concept discovery, concept-document tagging, and concept-hierarchy generation—to facilitate scientific knowledge exploration (see Figure 1). This is one of the core components in constructing the Microsoft Academic Graph (MAG), which enables a semantic search experience in the academic domain[1]. MAG is a scientific knowledge base and a heterogeneous graph with six types of academic entities: publication, author, institution, journal, conference, and field-of-study (i.e., *concept* or *FoS*). As of March 2018, it contains more than 170 million publications with over one billion paper citation relationships, and is the largest publicly available academic dataset to date[2].

To generate high-quality *concepts* with comprehensive coverage, we leverage Wikipedia articles as the source of concept discovery. Each Wikipedia article is an *entity* in a general knowledge base (*KB*). A *KB* entity associated with a Wikipedia article is referred to as a Wikipedia entity. We formulate concept discovery as a knowledge base type prediction problem (Neelakantan and Chang, 2015) and use graph link analysis to guide the process. In total, 228K academic concepts are identified from over five million English Wikipedia entities.

During the tagging stage, both textual information and graph structure are considered. The text from Wikipedia articles and papers' meta information (e.g., titles, keywords, and abstracts) are used as the *concept*'s and *publication*'s textual representations respectively. Graph structural information is leveraged by using text from a *publication*'s neighboring nodes in MAG (its citations, references, and publishing venue) as part of the *publication*'s representation with a discounting factor. We limit the search space for each publication to a constant range, reduce the complexity to $O(N)$ for scalability, where N is the number of publications. Close to one billion concept-publication pairs are established with associated confidence scores.

Together with the notion of subsumption (Sanderson and Croft, 1999), this confidence score is then used to construct a six-level directed acyclic graph (DAG) hierarchy with over 200K nodes and more than one million edges.

Our system is a deployed product with regular data refreshment and algorithm improvement. Key features of the system are summarized in Table 1. The system is updated weekly or monthly to include fresh content on the Web. Various document and language understanding techniques are experimented with and incorporated to incrementally improve the performance over time.

2 System Description

2.1 Concept Discovery

As top level disciplines are extremely important and highly visible to system end users, we man-

ually define 19 top level ("L0") disciplines (such as *physics*, *medicine*) and 294 second level ("L1") sub-domains (examples are *machine learning*, *algebra*) by referencing existing classification[3] and get their correspondent Wikipedia entities in a general in-house knowledge base (*KB*).

It is well understood that entity types in a general *KB* are limited and far from complete. Entities labeled with *FoS* type in *KB* are in the lower thousands and noisy for both in-house *KB* and latest Freebase dump[4]. The goal is to identify more *FoS* type entities from over 5 million English Wikipedia entities in an in-house KB. We formulate this task as a knowledge base type prediction problem, and focus on predicting only one specific type—*FoS*.

In addition to the above-mentioned "L0" and "L1" FoS, we manually review and identify over 2000 high-quality ones as initial seed FoS. We iterate a few rounds between a *graph link analysis* step for candidate exploration and an *entity type based filtering and enrichment* step for candidate fine-tuning based on *KB* types.

Graph link analysis: To drive the process of exploring new FoS candidates, we apply the intuition that if the majority of an entity's nearest neighbours are FoS, then it is highly likely an FoS as well. To calculate nearest neighbours, a distance measure between two Wikipedia entities is required. We use an effective and low-cost approach based on Wikipedia link analysis to compute the semantic closeness (Milne and Witten, 2008). We label a Wikipedia entity as an FoS candidate if there are more than K neighbours in its top N nearest ones are in a current FoS set. Empirically, N is set to 100 and K is in $[35, 45]$ range for best results.

Entity type based filtering and enrichment: The candidate set generated in the above step contains various types of entities, such as *person*, *event*, *protein*, *book topic*, etc.[5] Entities with obvious invalid types are eliminated (e.g. *person*) and entities with good types are further included (e.g. *protein*, such that all Wikipedia entities which have labeled type as *protein* are added).

The results of this step are used as the input for *graph link analysis* in the next iteration.

More than 228K FoS have been identified with this iterative approach, based on over 2000 initial seed FoS.

2.2 Tagging Concepts to Publications

We formulate the concept tagging as a multi-label classification problem; i.e. each publication could be tagged with multiple FoS as appropriate. In a naive approach, the complexity could reach $M \cdot N$ to exhaust all possible pairs, where M is 200K+ for FoS and N is close to 200M for publications. Such a naive solution is computationally expensive and wasteful, since most scientific publications cover no more than 20 FoS based on empirical observation.

We apply heuristics to cut candidate pairs aggressively to address the scalability challenge, to a level of 300–400 FoS per publication[6]. Graph structural information is incorporated in addition to textual information to improve the accuracy and coverage when limited or inadequate text of a *concept* or *publication* is accessible.

We first define **simple** *representing text* (or *SRT*) and **extended** *representing text* (or *ERT*). *SRT* is the text used to describe the academic entity itself. *ERT* is the extension of *SRT* and leverages the graph structural information to include textual information from its neighbouring nodes in MAG.

A publishing venue's full name (i.e. the journal name or the conference name) is its *SRT*. The first paragraph of a concept's Wikipedia article is used as its *SRT*. Textual meta data, such as title, keywords, and abstract is a publication's *SRT*.

We sample a subset of publications from a given venue and concatenate their *SRT*. This is used as this venue's *ERT*. For broad disciplines or domains (e.g. L0 and L1 FoS), Wikipedia text becomes too vague and general to best represent its academic meanings. We manually curate such concept-venue pairs and aggregate *ERT* of venues associated with a given concept to obtain the *ERT* for the concept. For example, *SRT* of a subset of papers from *ACL* are used to construct *ERT* for *ACL*, and subsequently be part of the *ERT* for *natural language processing* concept. A *publication*'s *ERT* includes *SRT* from its citations, references and *ERT* of its linked publishing *venue*.

---

[3] http://science-metrix.com/en/classification
[4] https://developers.google.com/freebase/
[5] Entity types are obtained from the in-house KB, which has higher type coverage compared with Freebase, details on how the in-house KB produces entity types is out-of-scope and not discussed in this paper.

[6] We include all L0s and L1s and FoS entities spotted in a publication's **extended** *representing text*, which is defined later in this section

We use h_s^p and h_e^p to denote the representation of a *publication* (*p*)'s *SRT* and *ERT*, h_s^v and h_e^v for a *venue* (*v*)'s *SRT* and *ERT*. Weight w is used to discount different neighbours' impact as appropriate. Equation 1 and 2 formally define publication *ERT* and venue *ERT* calculation.

$$h_e^p = h_s^p + \sum_{i \in Cit} w_i h_s^p(i) + \sum_{j \in Ref} w_j h_s^p(j) + w_v h_e^v \quad (1)$$

$$h_e^v = \sum_{i \in V} h_s^p(i) + h_s^v \quad (2)$$

Four types of features are extracted from the text: bag-of-words (BoW), bag-of-entities (BoE), embedding-of-words (EoW), and embedding-of-entities (EoE). These features are concatenated for the vector representation h used in Equation 1 and 2. The *confidence score* of a concept-publication pair is the cosine similarity between these vector representations.

We pre-train the word embeddings by using the skip-gram (Mikolov et al., 2013) on the academic corpus, with 13B words based on 130M titles and 80M abstracts from English scientific publications. The resulting model contains 250-dimensional vectors for 2 million words and phrases. We compare our model with pre-trained embeddings based on general text (such as Google News[7] and Wikipedia[8]) and observe that the model trained from academic corpus performs better with higher accuracy on the concept-tagging task with more than 10% margin.

Conceptually, the calculation of publication and venue's *ERT* is to leverage neighbours' information to represent itself. The MAG contains hundreds of millions of nodes with billions of edges, hence it is computationally prohibitive by optimizing the node latent vector and weights simultaneously. Therefore, in Equation 1 and 2, we initialize h_s^p and h_s^v based on textual feature vectors defined above and adopt empirical weight values to directly compute h_e^p and h_e^v to make it scalable.

After calculating the similarity for about 50 billion pairs, close to 1 billion are finally picked based on the threshold set by the confidence score.
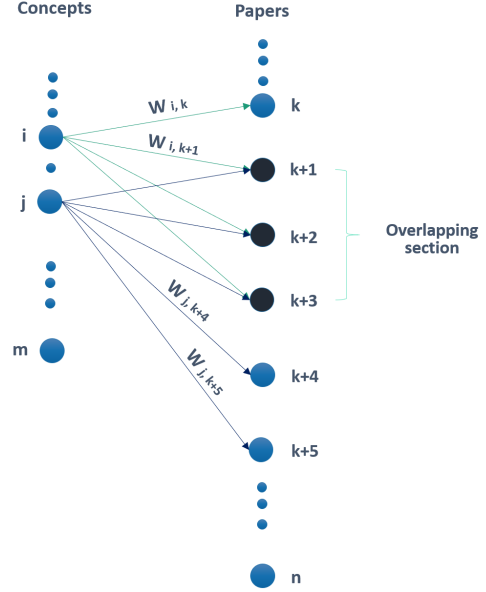
Figure 2: Extended subsumption for hierarchy generation.

2.3 Concept Hierarchy Building

In this subsection, we describe how to build a concept hierarchy based on concept-document tagging results. We extend Sanderson and Croft's early work (1999) which uses the notion of subsumption—a form of co-occurrence—to associate related terms. We say term x subsumes y if y occurs only in a subset of the documents that x occurs in. In the hierarchy, x is the parent of y. In reality, it is hard for y to be a strict subset of x. Sanderson and Croft's work relaxed the subsumption to 80% (e.g. $P(x|y) \geq 0.8, P(y|x) < 1$).

In our work, we extend the concept co-occurrence calculation weighted with the concept-document pair's confidence score from previous step. More formally, we define a *weighted relative coverage* score between two concepts i and j as below and illustrate in Figure 2.

$$RC(i,j) = \frac{\sum_{k \in (I \cap J)} w_{i,k}}{\sum_{k \in I} w_{i,k}} - \frac{\sum_{k \in (I \cap J)} w_{j,k}}{\sum_{k \in J} w_{j,k}} \quad (3)$$

Set I and J are documents tagged with concepts i and j respectively. $I \cap J$ is the overlapping set of documents that are tagged with both i and j. $w_{i,k}$ denotes the confidence score (or weights) between concept i and document k, which is the final *confidence score* in the previous concept-publication tagging stage. When $RC(i,j)$ is greater than a given positive threshold[9], i is the child of j. Since
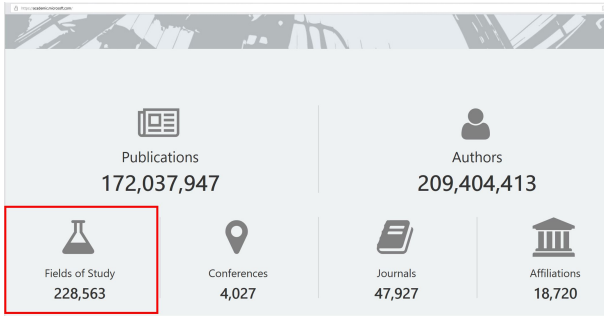
---

[9] It is usually in [0.2, 0.5] based on empirical observation.

Figure 3: Deployed system homepage at March 2018, with all six types of entities statistics: over 228K *fields-of-study*.

Step	Accuracy
1. Concept discovery	94.75%
2. Concept tagging	81.20%
3. Build hierarchy	78.00%

Table 2: Accuracy results for each step.

concepts[12] (*word embedding*, *artificial neural network*, *deep learning*, etc.) are shown on the right rail pane. Top tagged publications (without *word2vec* explicitly stated in their text) are recognized via graph structure information based on citation relationship.

this approach does not enforce single parent for any FoS, it results in a directed acyclic graph (DAG) hierarchy.

With the proposed model, we construct a six level FoS hierarchy (from L0 and L5) on over 200K concepts with more than 1M parent-child pairs. Due to the high visibility, high impact and small size, the hierarchical relationships between L0 and L1 are manually inspected and adjusted if necessary. The remaining L2 to L5 hierarchical structures are produced completely automatically by the extended subsumption model.

One limitation of subsumption-based models is the intransitiveness of parent-child relationships. This model also lacks a type-consistency check between parents and children. More discussions on such limitations with examples will be in evaluation section 3.2.

3 Deployment and Evaluation

3.1 Deployment

The work described in this paper has been deployed in the production system of Microsoft Academic Service[10]. Figure 3 shows the website homepage with entity statistics. The contents of MAG, including the full list of FoS, FoS hierarchy structure, and FoS tagging to papers, are accessible via API, website, and full graph dump from *Open Academic Society*[11].

Figure 4 shows the example for *word2vec* concept. Concept definition with linked Wikipedia page, its immediate parents (*machine learning*, *artificial intelligence*, *natural language processing*) in the hierarchical structure and its related

3.2 Evaluation

For this deployed system, we evaluate the accuracy on three steps (*concept discovery*, *concept tagging*, and *hierarchy building*) separately.

For each step, 500 data points are randomly sampled and divided into five groups with 100 data points each. On *concept discovery*, a data point is an FoS; on *concept tagging*, a data point is a concept-publication pair; and on *hierarchy building*, a data point is a parent-child pair between two concepts. For the first two steps, each 100-data-points group is assigned to one human judge. The concept hierarchy results are by nature more controversial and prone to individual subjective bias, hence we assign each group of data to three judges and use majority voting to decide final results.

The accuracy is calculated by counting positive labels in each 100-data-points group and averaging over 5 groups for each step. The overall accuracy is shown in Table 2 and some sampled hierarchical results are listed in Table 3.

Most hierarchy dissatisfaction is due to the intransitiveness and type-inconsistent limitations of the subsumption model. For example, most publications that discuss the *polycystic kidney disease* also mention *kidney*; however, for all publications that mentioned *kidney*, only a small subset would mention *polycystic kidney disease*. According to the subsumption model, *polycystic kidney disease* is the child of *kidney*. It is not legitimate for a *disease* as the child of an *organ*. Leveraging the entity type information to fine-tune hierarchy results is in our plan to improve the quality.

---

[10] https://academic.microsoft.com/
[11] https://www.openacademic.ai/oag/

[12] Details about how to generate related entities are out-of-scope and not included in this paper.
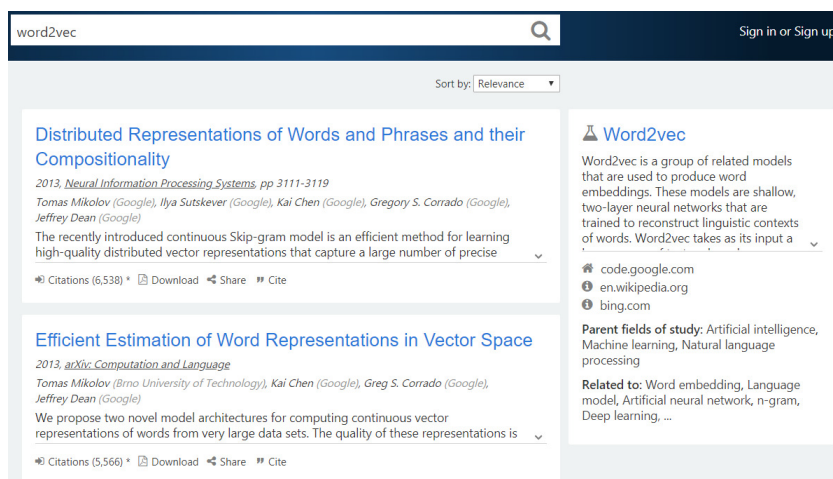
Figure 4: *Word2vec* example, with its parent *FoS*, related *FoS* and top tagged publications.

L5	L4	L3	L2	L1	L0
Convolutional Deep Belief Networks	Deep belief network	Deep learning	Artificial neural network	Machine learning	Computer Science
(Methionine synthase) reductase	Methionine synthase	Methionine	Amino acid	Biochemistry / Molecular biology	Chemistry / Biology
(glycogen-synthase-D) phosphatase	Phosphorylase kinase	Glycogen synthase	Glycogen	Biochemistry	Chemistry
	Fréchet distribution	Generalized extreme value distribution	Extreme value theory	Statistics	Mathematics
Hermite's problem	Hermite spline	Spline interpolation	Interpolation	Mathematical analysis	Mathematics

Table 3: Sample results for *FoS* hierarchy.

4 Conclusion

In this work, we demonstrated a Web-scale production system that enables an easy exploration of scientific knowledge. We designed a system with three modules: concept discovery, concept tagging to publications, and concept hierarchy construction. The system is able to cover latest scientific knowledge from the Web and allows fast iterations on new algorithms for document and language understanding.

The system shown in this paper builds the largest cross-domain scientific concept ontology published to date, and it is one of the core components in the construction of the Microsoft Academic Graph, which is a publicly available academic knowledge graph—a data asset with tremendous value that can be used for many tasks in domains like data mining, natural language understanding, science of science, and network science.

References

Yuxiao Dong, Hao Ma, Zhihong Shen, and Kuansan Wang. 2017. A century of science: Globalization of scientific collaborations, citations, and innovations. In *KDD*, pages 1437–1446.

Xueqing Liu, Yangqiu Song, Shixia Liu, and Haixun Wang. 2012. Automatic taxonomy construction from keywords. In *KDD*, pages 1433–1441.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

David Milne and Ian H. Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. *WIKIAI*, pages 25–30.

Arvind Neelakantan and Ming-Wei Chang. 2015. Inferring missing entity type instances for knowledge base completion: New dataset and methods. In *NAACL*, pages 515–525.

Mark Sanderson and W. Bruce Croft. 1999. Deriving concept hierarchies from text. In *SIGIR*, pages 206–213.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Paul Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *WWW*, pages 243–246.