A Web-scale system for scientific knowledge exploration

Lead Speaker: *Ramya Balasubramaniam (NLP Engineer, Chillwall.AI)*

Facilitators: *Ehsan Amjadian, Karim Khayrat*

Paper: *A Web-scale system for scientific knowledge exploration, Zhihong Shen, Hao Ma, Kuansan Wang*

Date: May 2, 2019

Important contribution of this paper

- Efficient exploration of Web-scale scientific knowledge.
- Organizing scientific publications into hierarchical concept structure.

Three essential requirements for this system

- To identify an exhaustive set of concepts that are covered by various publications.

- Associate these concepts to relevant set of publications.

- Build a six-level hierarchy of concepts with a subsumption-based model.

End Result: The system builds the most comprehensive cross-domain scientific concept ontology with more than 200 thousand concepts and over one million relationships.

Three essential components of this system

- Concept Discovery
- Concept-document tagging
- Concept-hierarchy generation

	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

End Result: MAG , Microsoft Academic Graph

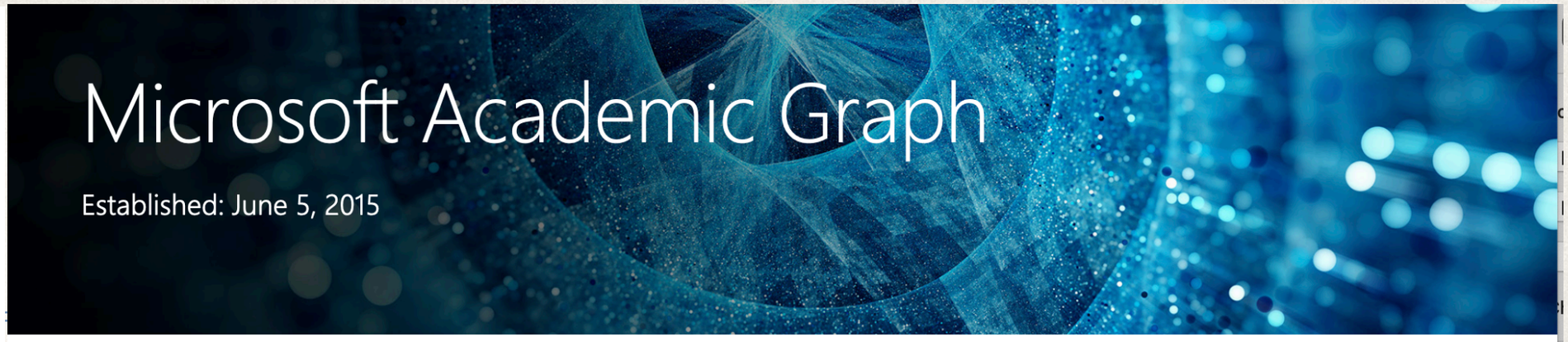
Microsoft Academic Graph

- Enables semantic search experience in the academic domain
- A scientific knowledge base and a heterogeneous graph with six types of academic entities: publication, author, institution, journal, conference and field-of-study
- As of March 2018, it contains more than 170 million publications with over one billion paper citation relationships
- Largest publicly available academic dataset to date.

Some online results for MAG

Kuansan Wang
Managing Director, MSR Outreach Academic Services

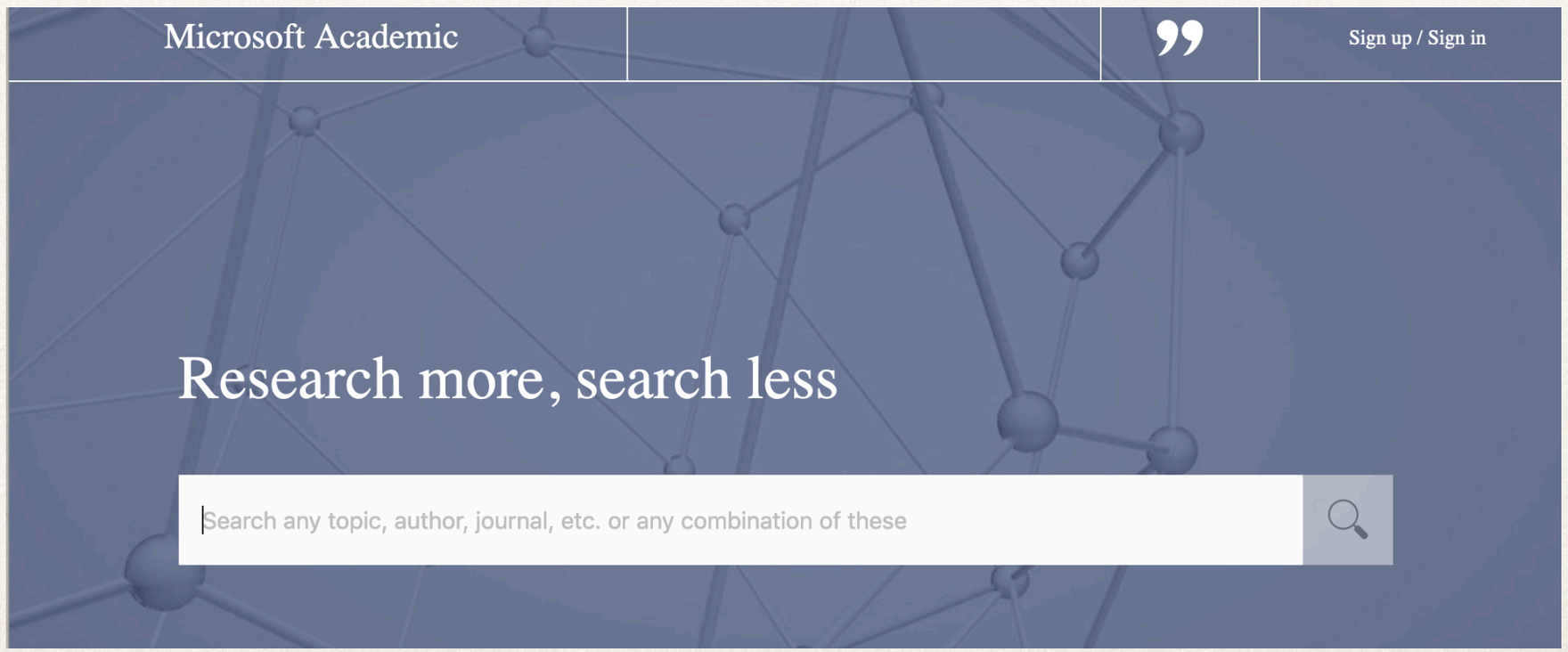
Original member of team



Overview | Publications | Microsoft Research blog | Projects

The Microsoft Academic Graph is a heterogeneous graph containing scientific publication records, citation relationships between those publications, as well as authors, institutions, journals, conferences, and fields of study. This graph is used to power experiences in Bing, Cortana, Word, and in Microsoft Academic. The graph is currently being updated on a weekly

Snapshot of Web portal describing MAG



Microsoft academic search tool

Comparison of Google Scholar & Microsoft Academic

- Link to Microsoft Academic
- Link to Google Scholar

Concept Discovery

- High-quality concepts were generated by leveraging Wikipedia articles. With each Wikipedia article taken as an entity in a knowledge base, Wikipedia entity.

- The problem of Concept Discovery is seen as Knowledge Base type prediction problem (Knowledge Base completion, KB is simply series of definitions describing an **is-a** relationship between entities such as Barack Obama was a US president, Tiger Woods is a sportsperson)

- In total, 228K academic concepts from over five million English Wikipedia Entities were identified.

Concept Discovery

- Top level concepts are highly visible to users, L0 level has 19 of them and L1 has 294 of them, extracted by referencing existing taxonomies such as Link to science metrix and their associated Wikipedia entities in the knowledge base.

- In the existing KBs (in-house Microsoft KB and Google Developer's Freebase dump, now obsolete) is limited and noisy. Identification of FoS type (Field-of-Study type) entities from over 5 million English Wikipedia entities.

- Three step process:

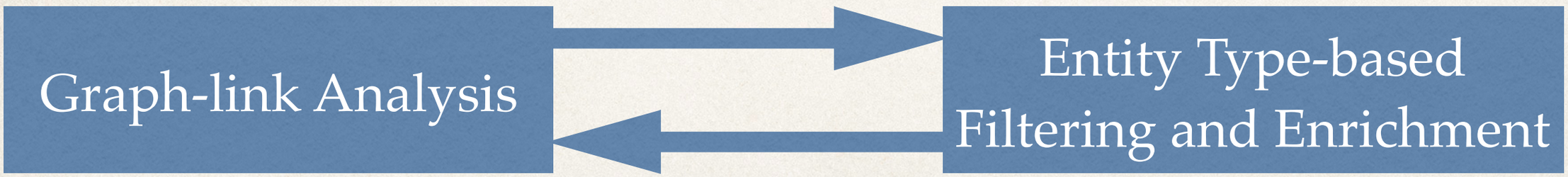
    1. **Initialization:** Initialize with 2000 high-quality manually selected entities as seed FoS

    2. **Graph Link Analysis:** Meant for candidate exploration (Included in iteration)

    3. **Entity Type-based Filtering and Enrichment :** Meant for fine-tuning candidates based on KB types. (Included in iteration)

Graph Link Analysis & Entity Type-Based Filtering and Enrichment

- We explore the KB looking for entities defined as FoS, then identify new ones using the intuition that if majority of neighbours of an entity in the KB graph are of FoS type, then the entity itself is of the FoS type.
- To calculate nearest neighbours a distance measure called WLM Wikipedia Link-based Measure is used, to calculate semantic closeness.
- A Wikipedia entity is labelled as FoS if out of Top N neighbours (typically 100) K neighbours (typically in the range [35,45]) are FoS.
- These identified candidates are verified using the original entity type associated with them, for.eg if the entity type is *person* then the candidate is eliminated and if the entity type is *protein* then the candidate is retained.



Concept-document Tagging

- Both textual and graph structure information is used for associating concepts and document.
- Textual Representation:
  1. Concept: Text of Wikipedia articles
  2. Publication: Paper's meta information (e.g. titles, keywords and abstracts)
- Graph Structure Information: Using Textual information from a publication's neighbouring nodes in MAG (its citations, references, and publishing venue), all of it included with a discounting factor (weight).
- Limit the search space for each publication to a constant range, reduce the complexity to O(N), N is number of publications.
- Close to one billion concept-publication pairs were established.

Concept-document Tagging

SRT or Simple Representing Text is the text used to describe the the academic entity itself. So, for:

1. Publication venue: SRT is just publishing venue name

2. Wikipedia article: First Paragraph of a concept's Wikipedia article.

3. Publication: Meta data such as Title, keywords, and abstract

ERT or Extended Representing Text is the extension of SRT:

1. Publication venue: A subset of publications from a given venue and concatenate their SRT.

2. Wikipedia article: For broad disciplines (in L0 and L1 level concepts) Wikipedia text is too vague, so manual curation of concept-venue pairs and then aggregate ERT of venues associated with the concept.

3. Publication: SRT of its citation, references and ERT of its linked publishing venue.

Concept-document Tagging

If h_s^p and h_e^p are SRT and ERT of publication, h_s^v and h_e^v are SRT and ERT of venues then

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

$$h_e^v = \Sigma_{i \in V} h_s^p(i) + h_s^v \ldots . (2)$$

Concept-document Tagging

- 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 vector representation *h* used in Eq. 1 and 2.
- Confidence score of a concept-publication pair is cosine similarity between these vector representations.
- They pre-trained word embeddings by using skip-gram on academic corpus, with 13 B words based on 130M titles and 80M abstracts from English scientific publications. Resulting model contains 250-dimensional vectors for 2 million words and phrases.
- Since, MAG contains hundreds of millions of nodes, its computationally prohibitive to optimize node latent vectors and weights simultaneously. So, the weights are empirically adapted.
- After calculating similarity for 50 billion pairs, close to one billion are picked based on a threshold set on the confidence score.

Concept Hierarchy Building

- The most important notion used here is that of subsumption, a form of co-occurrence , to associate related terms. x subsumes y if y occurs only in a subset of documents that x occurs in. But this need not be strictly true.

- More formally, the authors define a term called *weighted relative coverage score* between two concepts i and j where I and J are the set of documents tagged to concepts i and j.

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

  - So for concept i and j, $I \cap J$ are the documents that belong to concept i and concept j. $w_{i,k}$ is the weight associated with concept i and document k. If this score is greater than a positive threshold, then concept i is the child of concept j. This results in the formation of a DAG, since a particular child could have multiple parents.

A sample result for FoS hierarchy

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.

- Six FoS hierarchy (from L0 to L5) on over 200K concepts with more than 1 M parent-child pairs.
- Due to high visibility, high impact and small size, hierarchical relationships between L0 and L1 are manually inspected and adjusted if necessary. For the hierarchical relationships between concepts from the L2 to L5 levels a completely automated approach as described earlier is used.
- Two drawbacks: Intransitivity of parent and child relationship, lack of type-consistency check between child and parent.

Evaluation Process

- For this deployed system, all the three steps were evaluated for accuracy 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 FoS.
  - On concept tagging, a data point is concept-publication pair.
  - On hierarchy building, a data point is a parent-child pair between two concepts.
- For Concept discovery and Concept tagging , each of the 100-data-points group is assigned to a human judge.
- For concept hierarchy results are by nature more controversial and prone to individual subjective bias, each group is assigned to three judges and majority voting is used to decide final results.
- Accuracy is simply the proportion of positive labels averaged over all the five groups.

Discussion Points

- Is Intransitiveness of the parent-child relationship really a limitation?

- How are the weights or discounting factors associated with ERT vectors empirically calculated?