Probase: a Universal Knowledge Base for Semantic Search

Zhongyuan Wang, Jiuming Huang, Hongsong Li, Bin Liu, Bin Shao, Haixun Wang,
Jingjing Wang, Yue Wang, Wentao Wu, Jing Xiao, Kenny Q. Zhu

Microsoft Research Asia

ABSTRACT

We demonstrate a prototype system that showcases the power of using a knowledge base (Probase) for search. The goal of Probase is to enable common sense computing, and its foundation is a universal, probabilistic ontology that is more comprehensive than any of the existing ontologies. Currently, it contains over 2.7 million concepts harnessed automatically from a corpus of 1.68 billion web pages. Unlike traditional knowledge bases that treat knowledge as black and white, it supports probabilistic interpretations of the information it contains. The probabilistic nature also enables it to incorporate heterogeneous information in a natural way. Besides the system, we also demonstrate two applications, i) semantic web search and ii) understanding and searching web tables, that are built on top of the Probase framework. They indicate that a little common sense goes a long way: machines can be made more intelligent if it has access to the right knowledge.

1. THE KNOWLEDGE BASE

We demonstrate Probase, an ongoing project that focuses on knowledge acquisition and knowledge serving. In the age of information explosion, there is a pressing need to enable machines to better understand natural human language. The Probase project shows that such understanding might be made possible if we can harvest "general knowledge" or "common sense" from the enormous amount of data that is available to us, and inject them into computing.

The question is then, what is general knowledge or common sense, and can machines grasp common sense? It is certainly difficult for machines to master the entire body of general knowledge. However, it is still possible to give machines certain common sense, and it turns out that a little common sense goes a long way. For example, for human beings, when we see "25 Oct 1881", we recognize it as a date, but the majority of us do not know what the date is for. However, if we had a little more context, say it is embedded in the text "Pablo Picasso, 25 Oct 1881, Spain", we would have guessed (correctly) that it is Pablo Picasso's birthday. Human beings are able to do this because we possess certain common sense, and in this case, "one of the most important dates associated with a person is his birthday." Take natural language processing as another example. Humans do not find sentences such as "animals other than dogs such as cats" ambiguous, but machine parsing can lead to two possible understandings: "cats are animals" or "cats are dogs." Common sense tells us that cats cannot be dogs, which renders the second parsing mostly improbable. It turns out that all the common sense needed in the above two cases are about concepts (e.g., persons and animals), instances (e.g., Pablo Picasso, cats, and dogs), attributes (e.g., birthday), and values (e.g., 25 Oct 1881). It is not impossible to build a knowledge base that encode such common sense.

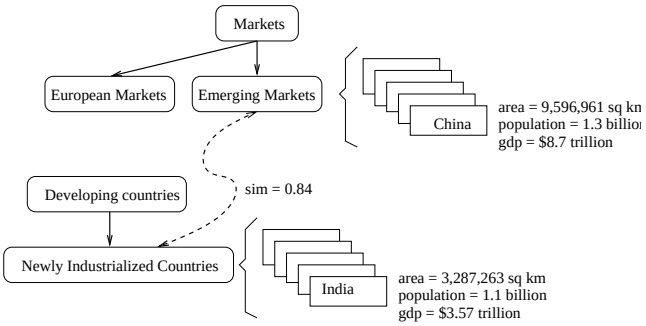

Figure 1: A Snippet of Probase's Core Taxonomy.

We build Probase for this purpose. Figure 1 shows a snippet of Probase, which contains concepts (e.g., Newly Industrialized Countries), instances (e.g., China), attributes and values (population = 1.3 billion), and relationships (e.g., similarity). But Probase is much more than a traditional ontology/taxonomy. It is unique because of its extremely rich conceptual space, and its probabilistic nature that allows it to incorporate other data. Probase's core taxonomy alone contains about 2.7 million concepts harnessed from a corpus of 1.68 billion web pages and 2 years' worth of search log. Figure 2 shows the popularity (# of occurrences in the corpus) distribution of the 2.7 million concepts. In contrast, the well known Freebase [4] taxonomy, which is built by community efforts, contains no more than 2,000 concepts, and Cyc [5], after 25 years of continuing improvement by human experts, has about 120,000 classes (concepts). As we can see in Figure 2, besides popular concepts such as "countries" and "companies", which are included by almost every general purpose taxonomy, Probase has millions of concepts such as "renewable energy technologies", "celebrity wedding dress designers" and "basic watercolor techniques", which cannot be found in Freebase or Cyc. However, not only are they concrete and meaningful, they are useful in interpreting human communications. Since Probase's 2.7 million concepts are har-

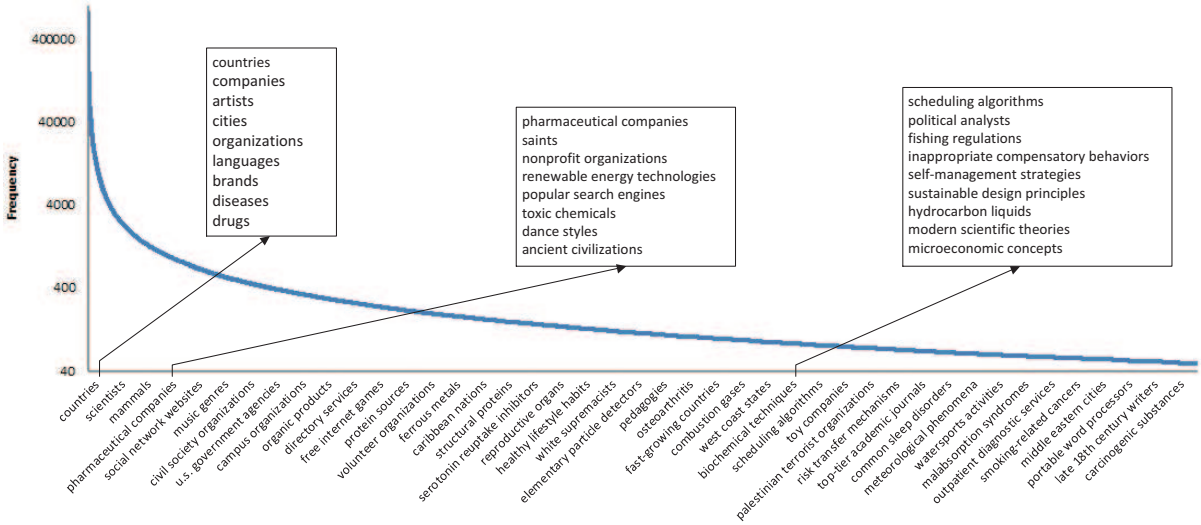

Figure 2: Distribution of the 2.7 million concepts in Probase

nessed from a corpus of 1.68 billion Web pages authored by millions of people, probably it already includes most, if not all, concepts of worldly facts that human beings have in their mind. With such a rich concept space, Probase has much better chance to understand writings (including keyword queries or natural language texts) created by human beings. Indeed, we studied 2 years' worth of Microsoft's Bing search log, and found that 85% of the searches contain concepts and/or instances that exist in Probase. It means Probase can be a powerful tool to interpret user intention behind search [8, 6].

Beyond the core taxonomy, Probase is able to incorporate data from heterogeneous sources by first trying to understand the data using the knowledge in its core taxonomy. The reason that Probase is able to accumulate a large amount of data is because of its probabilistic nature. We do not regard the data in Probase as facts, instead, we regard them as claims or beliefs associated with probabilities that model their plausibility, ambiguity, and other characteristics. Furthermore, we regard external data, such as the web, search engine log, dictionaries and encyclopedias, *etc.*, as evidences that can add to or modify the claims and beliefs in Probase.

2. THE ARCHITECTURE

Figure 3 presents the architecture of Probase. It has the following components.

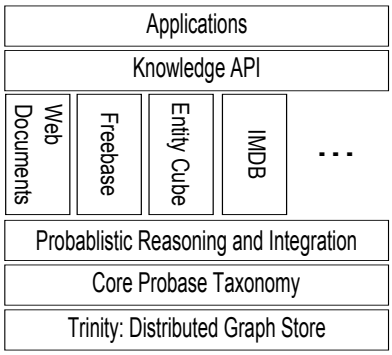

Figure 3: The Probase Architecture.

- The foundation of the Probase framework is a distributed hypergraph store called Trinity [2]. Trinity supports online parallel query processing of massive hypergraphs. The topology of the hypergraph is memory based, and can be hosted on one machine or thousands of machines.

- The core Probase taxonomy [9] is constructed using information obtained from a large web corpus of 1.68 billion pages as well as 2 years' worth of search log data.

- Probase's probabilistic reasoning and integration layer enables it to incorporate data of varied quality from heterogeneous sources, including Freebase [4], Wikipedia [3], *etc.* We are planning to incorporate data in vertical domains as well, including EntityCube [1], location data, IMDB movies, Amazon products, *etc.*

- On top of Probase, we support various applications [9, 7, 6, 8]. We provide a set of APIs which enables applications to access the probabilistic knowledge in Probase.

3. ABOUT THE DEMO

Our demonstration consists of the following parts.

Taxonomy Construction and Browsing. We demonstrate the iterative process that builds the core taxonomy: in each iteration we collect information that Probase can understand, and then we convert the information into new knowledge in Probase, which enables us to understand more information in the next iteration. The taxonomy browser (Fig. 4) allows users to explore the core Probase taxonomy with its 2.7 million classes and over 16 million instances. The browser provides a search interface for concepts, and shows a concept's isA hierarchy, its instances (entities), and its similar concepts.

Semantic Search on the Web. Since more than 85% of searches contain concepts and/or instances that can be found in Probase, Probase has a good advantage to interpret the intention of the user.

Consider the following search queries: i) *ACM fellows working on semantic web*; ii) *database conferences in asian cities*; iii)

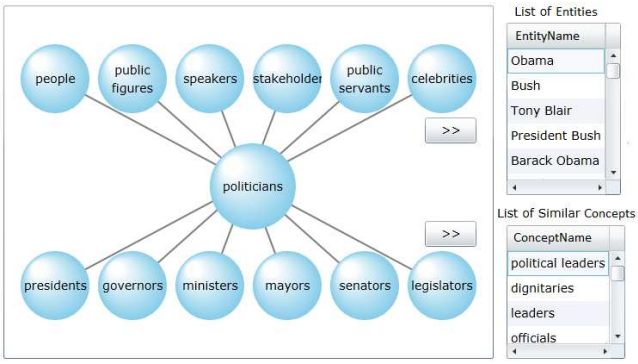

Figure 4: Browsing Probase's taxonomy

highest mountains in US; and iv) *winter vacation destinations except florida*. The user intention of each of the above queries is clear. However, current search engines fail to deliver good results. The reason is that keyword based search engines cannot interpret the concepts in the search. Instead, they find exact, word-for-word matches for phrases such as "database conferences", "asian cities", "ACM fellows", "mountains in US". Furthermore, they do not know that in order to find the "highest mountain", all we need to do is to apply the max aggregate on the *elevation* or *altitude* attribute of the mountain concept, and that "except florida" means the other 49 states in the US.

Fig. 5 shows our (complementary) search results and Bing's results for query *winter vacation destination except florida*. While Bing's results contain the keyword "Florida" in every result, our system correctly interprets the query and returns winter destinations other than Florida.

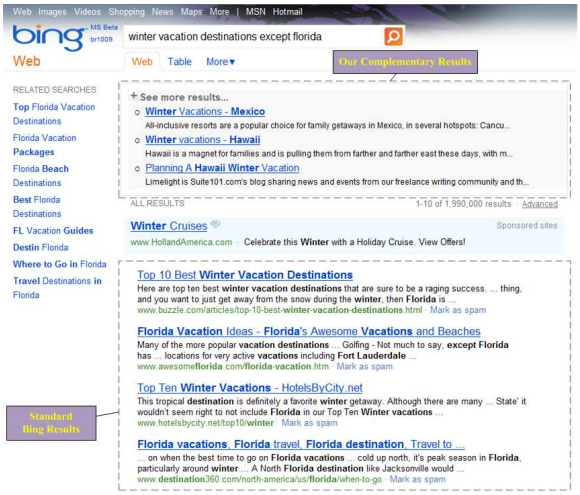

Figure 5: Semantic search results compared with Bing results

Understanding and Searching Tables. We use Probase to unlock the information in tables on the web, and the information, once understood, is used to enrich Probase. The reason we focus on tables on the Web is two-fold. First, there are billions of tables on the Web, and they contain much valuable information. Second, tables are relatively well structured, which means they are easier to understand than text in natural languages.

Fig. 6 shows the search interface for tables. Given query "politicians birthday", Probase returns tables for U.S. Vice President, Senators, *etc.*, with columns including Birthday or Date of birth, *etc.* The first step of understanding tables lies in finding the schema of the table, and this is made possible by repeatedly invoking two Probase knowledge APIs: `findConceptForAttributes()` and `findConceptForEntities()`. Each function returns a list of concepts each associated probabilities. We then pinpoint the most likely concept by maximizing the likelihood.

Figure 6: Table search results

4. SYSTEM & DATA AVAILABILITY

The Probase taxonomy will be made available to the public in the near future. Currently, more information about Probase, including its framework, the semantic search and the table search demo, as well as a small excerpt of the Probase taxonomy, can be found at http://research.microsoft.com/probase/.

5. REFERENCES

[1] Entitycube. `http://entitycube.research.microsoft.com/`.

[2] Trinity: A distributed hypergraph store. `http://research.microsoft.com/trinity/`.

[3] Wikipedia. `http://www.wikipedia.org`.

[4] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, 2008.

[5] D. B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley, 1989.

[6] J. Wang, B. Shao, H. Wang, and K. Q. Zhu. Understanding tables on the web. Technical report, Microsoft Research, 2010.

[7] P. Wang, H. Li, H. Wang, and K. Q. Zhu. Taxonomy assisted massive relationship extraction from the web. 2010.

[8] Y. Wang, H. Li, H. Wang, and K. Q. Zhu. Toward topic search on the web. Technical report, Microsoft Research, 2010.

[9] W. Wu, H. Li, H. Wang, and K. Q. Zhu. Inferring a universal probabilistic taxonomy from the web. Technical report, Microsoft Research, 2010.