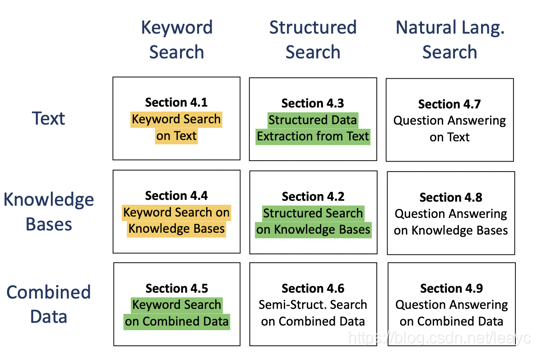
# Semantic searching

## Introduction

Semantic search is when a whole lot of resources are used in order to perform a search, rather than just keywords.

Semantic Search was formally proposed in 2007~2009. Corresponding to it is lexical search, which uses one-to-one correspondence between literal values or string similarity for resource recall. The disadvantages of this type of method are obvious-it cannot handle the same name, aliases, and complicated situations.

In order to sort out relevant work in a more detailed way,Bast et al. divided semantic Search into 9 situations according to two dimensions of data and Search.



Specifically, Bast believes that the most basic data can be divided into two types: text and knowledge base. With the development of the semantic web/knowledge graph, there has also been a combination of Semantic marked text, the most typical such as html5; search can usually be divided into keyword search, structured search and natural language search. Pairwise combination, there are nine situations on the picture.

In addition, there are also methods such as facet search and representation search. Faceted search is actually quite common:



Click the limited tag layer by layer to get the corresponding result in the shopping or life service website, which is the most typical faceted search. As for the representation search, it generally refers to the low-dimensional embedding of the node relationship in the knowledge graph before processing.

Semantic search takes into consideration user intent and user data.

This is perhaps the most important thing about semantic search. The very best results aren’t those that contain lots of keywords, an optimized H1, and a well-crafted title tag.

Instead, the most relevant page is one that aligns with the user’s intent. Based on the aggregate data of millions of searches, Google’s machine-learning algorithm has trained itself to interpret what you really want.

## Related work

Cognitive Ontology Enrichment for Semantic Information Retrieval for semantic search system is obtained by first converting HTML to XML, next converting XML to OWL Mapping and performing search using SPRQL [3]. Semantic search engines likes woogle, falcon, SWSE have structured data that handles data in OWL or RDF format only. Moreover, XML takes in accord the Syntactic level but not the reasoning [4]. Employing SWSE(Semantic Web Search Engine) for Searching and Browsing Linked Data: [5]. Retrieval Comparison is carried out by relying upon Precision and Recall for information. It’s depicted that semantic precision is 0.72 and syntactic precision is 0.42[4].Varying work survey is generated concerning this domain. Based on the precision and recall result, the evaluation of search engine’s performance is performed.

There are search engines for various open datasets. Kaggle is the world’s largest community of data scientists and ma-chine learners (https://www.kaggle.com/). Kaggle datasets have no or very simple dataset description. Dataset search is by tags indicating the topics, classifications or data types provided by dataset owners. Kaggle dataset listing is ranked by hotness (interestingness and recency) by default another methods of sorting like Most Votes, New, Recently Active, and Updated. CKAN (https://ckan.org/) is the open source data portal platform widely used for implementing open data catalogue systems, with more than 100 data hubs internationally , including government data catalogues for UK (data.gov.uk), USA (catalog.data.gov), EuropeanUnion (publicdata.eu) and others. CKAN’s dataset search iskeyword-based and provides full-text search and facet search(e.g., by tags). The recently announced Google dataset search is based on structured data that data publishers embedded on the web pages about the datasets using the schema.org Dataset markup or DCAT. All these dataset search engines provide only keyword-based search.

## Problem Defination

In the Semantic searching part, we will implement an application that support users to use semantic words to search entities in the knowledge graph. Given a word **w**, we should get the entity **e** which is correspond to the word w, i.e., the entity **e** is describing the word **w**. However, it is impossible to find the definitely corresponding one, so we should list the top-10 most similar entities set S={e1,e2...e10} as the final result.

**Approach**

It is very obviously that this problem is very similar to the Knowledge Integration part because we just need to calculate the similarity of two entities. So we can calculate the similarity between input word embedding and the embedding of the label of the entities, and then sort the similarities we calculate. Finally, we set the top-10 entities to be our final result.

**Experiment**

We have the approach to get the word embedding in the Knowledge Extraction part, so we just use the BERT pre-trained model to get the input word embedding and the entity label embedding. Meanwhile, we use cos distant as the semantic distance between words.

## Reference

[1] Li, Hang, and Jun Xu. “Semantic matching in search.” Foundations and Trends® in Information Retrieval 7.5 (2014): 343-469.  
 [2] Bast, Hannah, Björn Buchhold, and Elmar Haussmann. “Semantic search on text and knowledge bases.” Foundations and Trends® in Information Retrieval 10.2-3 (2016): 119-271.

[3] G. Nagarajan , K.K. Thyagharajan, “cognitiveOntology Enrichment For Semantic Information Retrieval”, National Journal Of System And Information Technology, Vol 5, No 2 , 2012,pp 168-179.

[4]Ranjna Jain, Neelam Duhan and A.K. Sharma, "Comparative Study on Semantic Search Engines", International Journal of Computer Applications (0975 – 8887), Volume 131 – No.14, December 2015.

[5]Aidan Hogan, Andreas Harth , J• urgen Umbrich , Sheila Kinsella , Axel Polleres , Stefan Decker, “Searching and Browsing Linked Data with SWSE: the Semantic Web Search Engine”, Journal of Web Semantics, Vol 9 no.4 , 2011.

[6] G. Antoniou and F. van Harmelen, A semantic web primer. The MIT Press, 2004.  
 [7] S. Sitthisarn, L. Lau, and P. M. Dew, “A survey of semantic keyword search approaches,” Int. J. of Applied Computer Technology and Information Systems, vol. 2, no. 2, 2013.