SetSearch: A Set-Oriented and Entity-Aware Search System for Biomedical Literature

Presenter: Jiaming Shen

Advisor: Jiawei Han

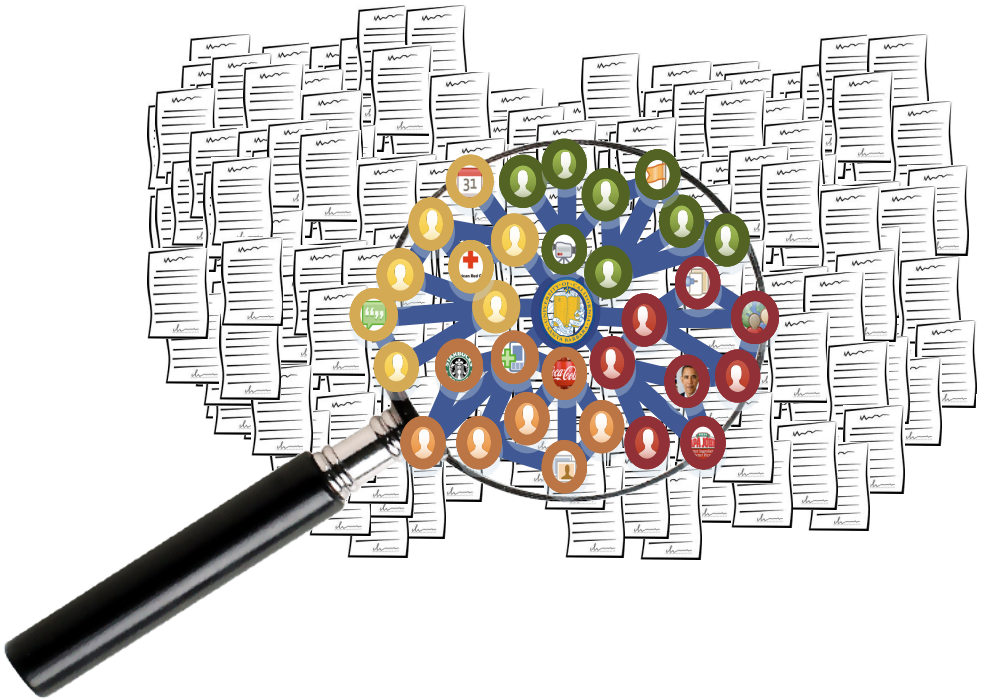
Jan. 30, 2017

Outline

- Introduction 
- Related works
- Proposed method
- System Implementation
- Results Analysis
- Future works

Introduction

- Literature search is fundamental for biomedical research!

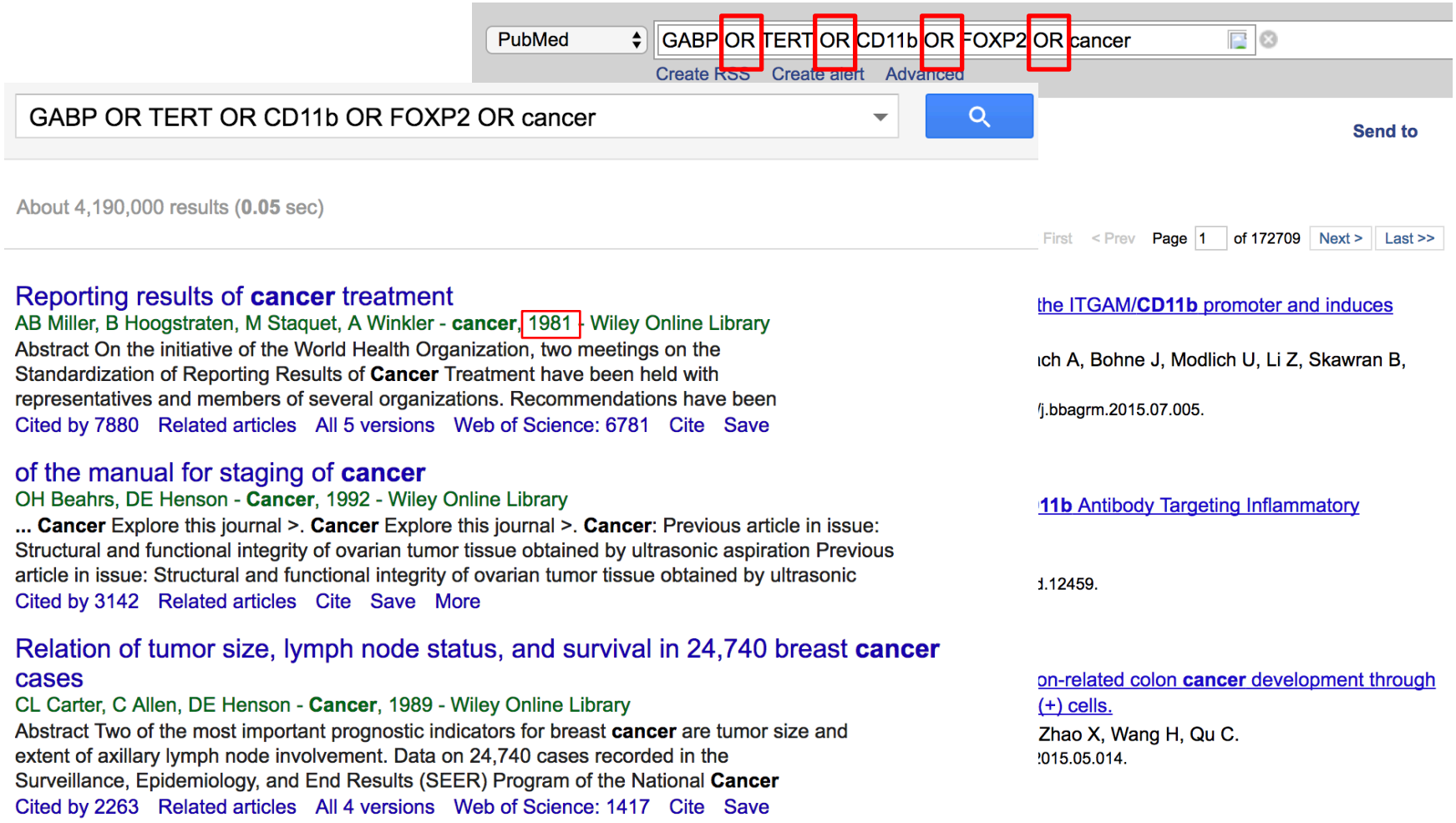


Google Scholar

PubMed

Introduction

- Existing search engines do not perform well!

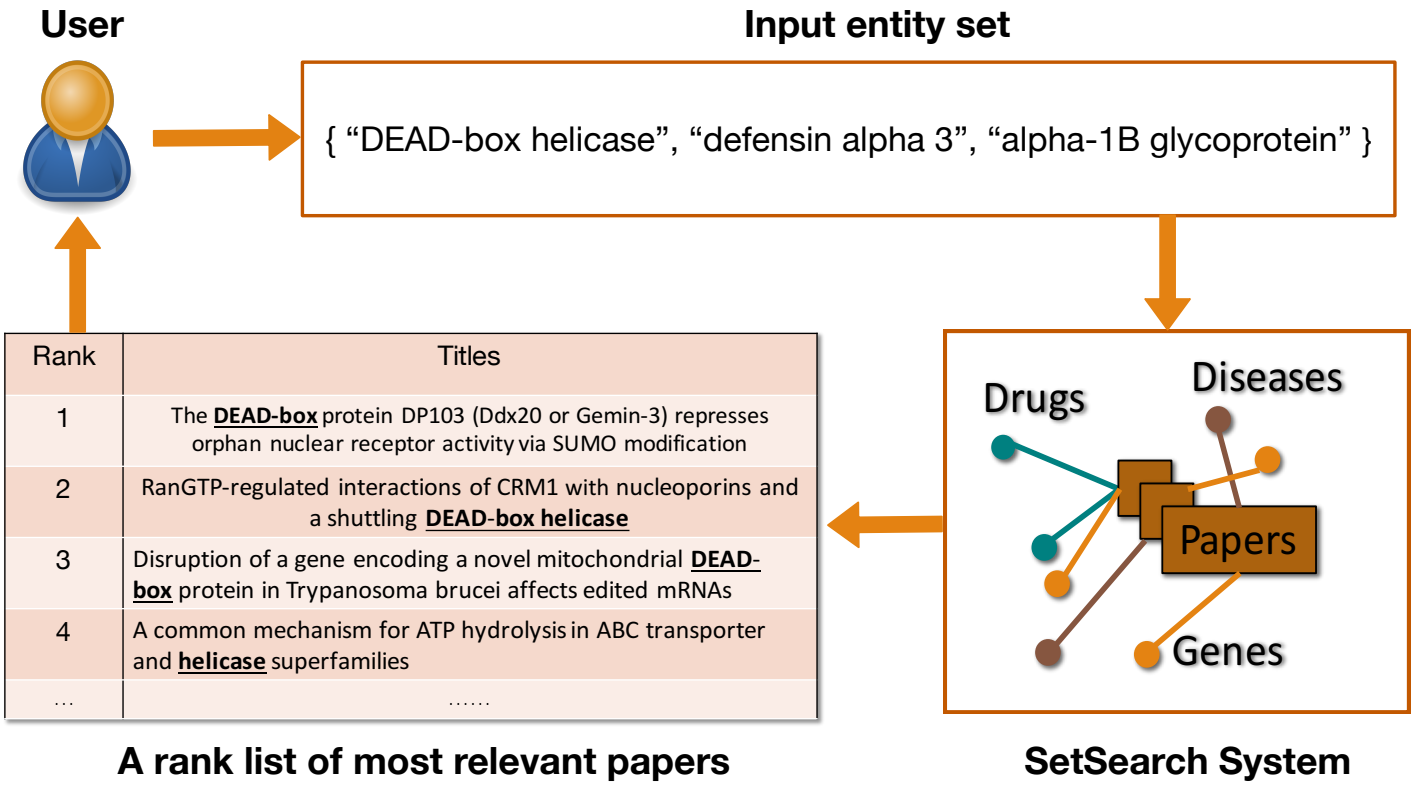


Motivation

We need to do better for entity set query !

Problem Definition

- Given **a set of entity** (gene, disease, chemical, etc), output a ranked list of papers **most relevant** to the whole set of entities.



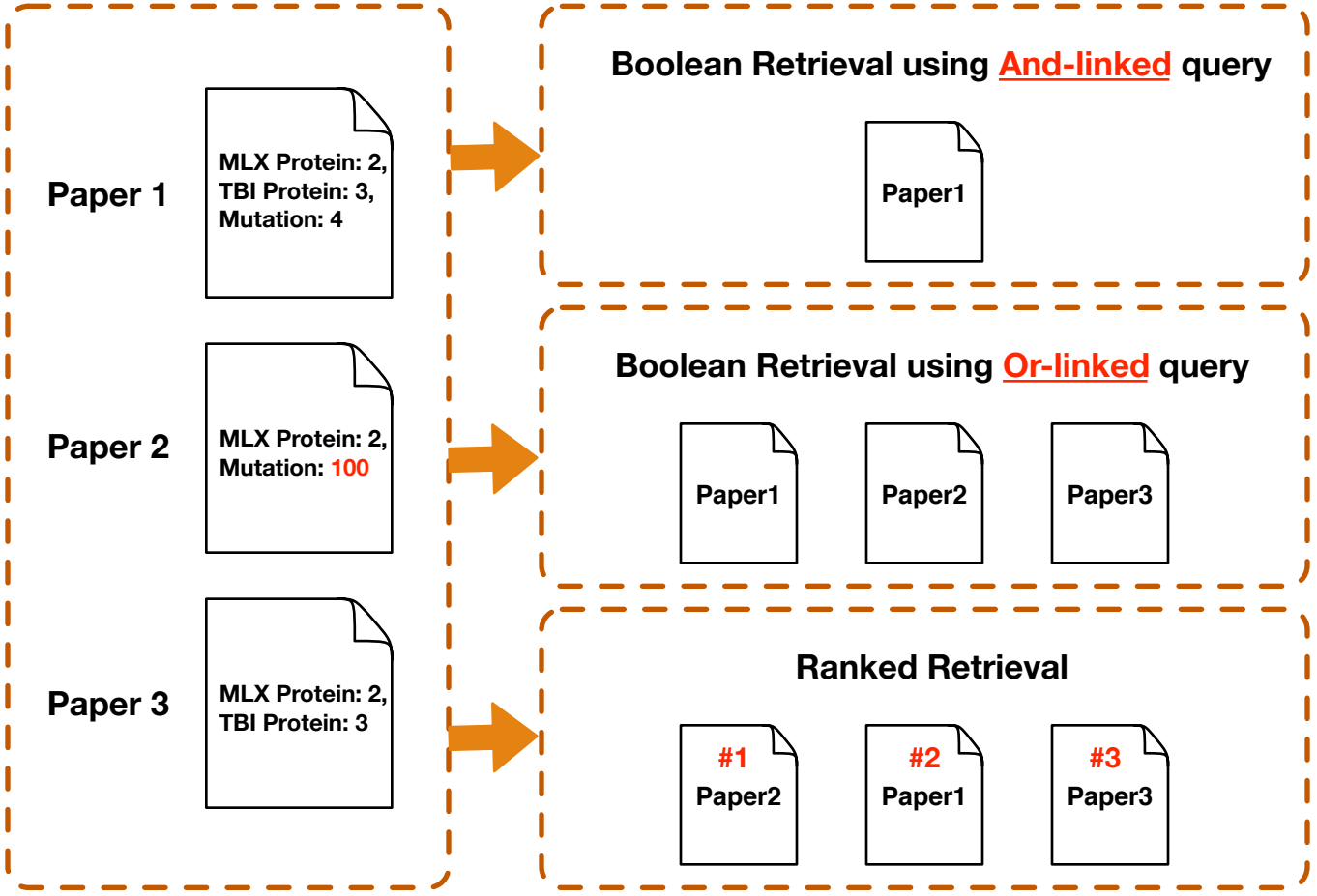
Outline

- Introduction
- Related works 
- Proposed method
- System Implementation
- Results Analysis
- Future works

Related works

- Boolean Retrieval:
  - One paper is either fully related (1) or completely not (0).
    - Or-linked query: retrieve a paper containing **any** gene name.
    - And-linked query: retrieve a paper containing **all** gene names.
- Ranked Retrieval:
  - Return a ranked list of paper based on **independent words**
  - Including TF-IDF, BM25, etc.

Related works



Query = {MLX Protein, TBI Protein, Mutation}

Related works

- Drawbacks of current method
  - And-linked query
    - too restrictive, cannot get enough papers
  - Or-linked query & ranked retrieval:
    - too biased toward very popular gene names
  - A gap between And-linked & Or-linked query

Number of gene input	2	3	4	5
Avg. number of papers retrieved by **And-linked query** (out of 13,839 full papers)	0.487	0.001	0.0	0.0
Avg. number of papers retrieved by **Or-linked query** (out of 13,839 full papers)	290.1	423.3	507.1	635.4

Outline

- Introduction
- Related works
- Proposed method 
- System Implementation
- Results Analysis
- Future works

Proposed method - Overview

- Design Goals
  - Incorporate entity information such as entity type.
    - E.g., if we know "GABP" is gene, and "prostate cancer" is a disease, how can we use these information?
  - Consider entity set property.
    - E.g., for query "GABP, TERT, cancer", how to treat it as an entity set instead of each individual terms?
- Three components
  - Entity Language Model
  - Query Model
  - Graph covering ranking principle

Entity Language Model

- We assume the generative process of each document as follows:

For each token t in document d_i: Gene, disease

(a) we generate the type of that token $\phi(t)$ based on $P(\phi(t)|\alpha_{d_i})$,

(b) we generate the token t based on $P(t|\theta_{\phi(t)|d_i})$.

"GABP", "cancer"

Parameters

- using data statistics, we can learn model parameters as follows:

$$P(\phi(t)|\alpha_{d_i}) = \frac{n_{\phi(t),d} + \mu_\alpha \frac{n_{\phi(t),D}}{L_D}}{L_d + \mu_\alpha}, \quad P(t|\theta_{\phi(t)|d_i}) = \frac{n_{t,d} + \mu_\theta \frac{n_{t,D}}{n_{\phi(t),D}}}{n_{\phi(t),d} + \mu_\theta},$$

Table 1: Parameter Explanations

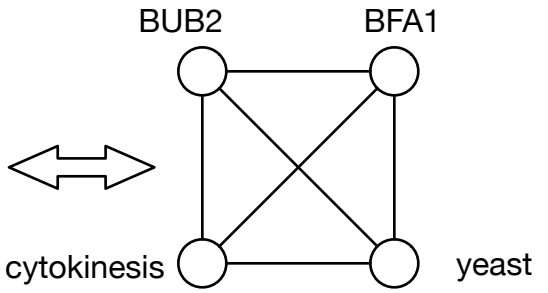
Name	Description
L_x	length of text sequence x, x can be document d_i, query q or corpus D
$n_{t,x}$	number of token t in text sequence x
$n_{\phi(t),x}$	number of tokens with type $\phi(t)$ in text sequence x
α_x	type distribution of text sequence x
$\theta_{\phi(t)\|x}$	token distribution of type $\phi(t)$ in text sequence x
μ_α, μ_θ	hyper-parameters for Dirichlet smoothing

Using Dirichlet smoothing

Query Model

- We use a graph to represent a query.

Query = {BUB2, BFA1, cytokinesis, yeast}



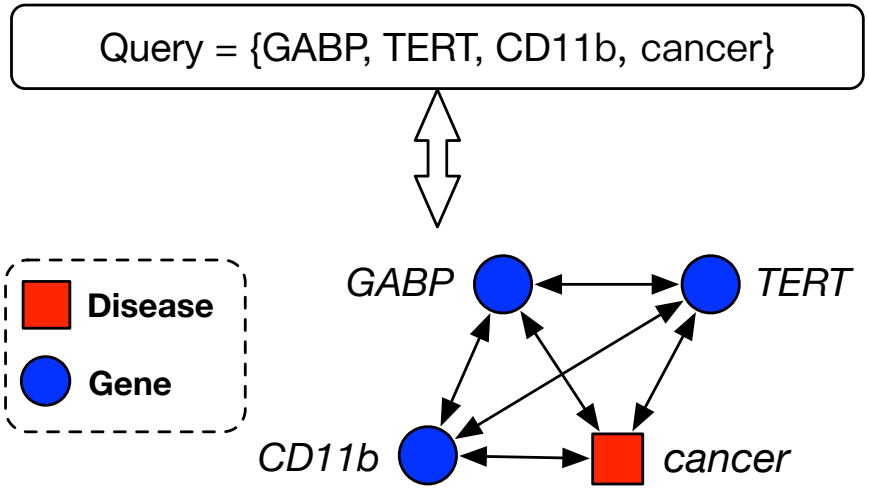
Query Graph Representation

Node: a query term
Edge: a latent relation

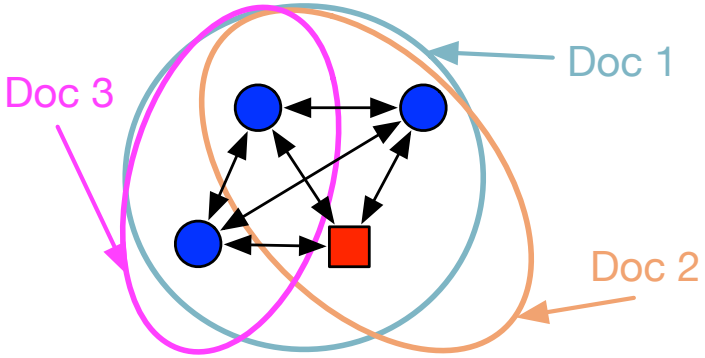
- Exist many other graph structures:
  - Complete graph (all relations are important)
  - Graph with only nodes (no relation is important)
  - Hyper-graph (each node means an synonym set)

Graph Covering Ranking Principle

- Ranking Principle: a document is ranked higher if it covers more latent query information.



Query Graph:
Node: a query term; Edge: a latent relation



Ranking order:
Doc 1 > Doc 2 > Doc 3

Ranking Principle:
A document is ranked higher if
it covers more latent query information.

Graph Covering Ranking Principle

- Technically, a document is ranked by the probability that it is generated by a query, which is further defined on the graph that is covered by the document.

We treat the graph as a **Markov network** and define P(M_di, q) based on it.

$$P(M_{d_i}|q) = \frac{P(M_{d_i}, q)}{P(q)}$$

$$\stackrel{\text{rank}}{=} \log P(M_{d_i}, q)$$

$$\stackrel{\text{rank}}{=} \sum_{c \in G_{q|d_i}} \log \psi(c)$$

$$= \sum_{c \in G_{q|d_i}} f(c).$$

The probability that a document is generated by the query

$$P(q, d_i) = \frac{1}{Z} \prod_{c \in G_{q|d_i}} \psi(c),$$

We further use **pairwise factorization** of a Markov network and thus each clique c is either **a node** or **an edge**.

Graph Covering Ranking Principle

We treat the graph as a **Markov network** and define P(M_di, q) based on it.

$$P(M_{d_i}|q) = \frac{P(M_{d_i}, q)}{P(q)}$$

$$\stackrel{\text{rank}}{=} \log P(M_{d_i}, q)$$

$$\stackrel{\text{rank}}{=} \sum_{c \in G_{q|d_i}} \log \psi(c)$$

$$= \sum_{c \in G_{q|d_i}} f(c).$$

The probability that a document is generated by the query

$$P(q, d_i) = \frac{1}{Z} \prod_{c \in G_{q|d_i}} \psi(c),$$

We further use **pairwise factorization** of a Markov network and thus each clique c is either **a node** or **an edge**.

For a node: $f(t) = \lambda_t \cdot \log P(t|M_{d_i}),$

Term weight

For an edge: $f(t_a, t_b) = \lambda_{t_a,t_b} \cdot \log P(t_a, t_b|M_{d_i}),$

Relation weight

Practical Issues

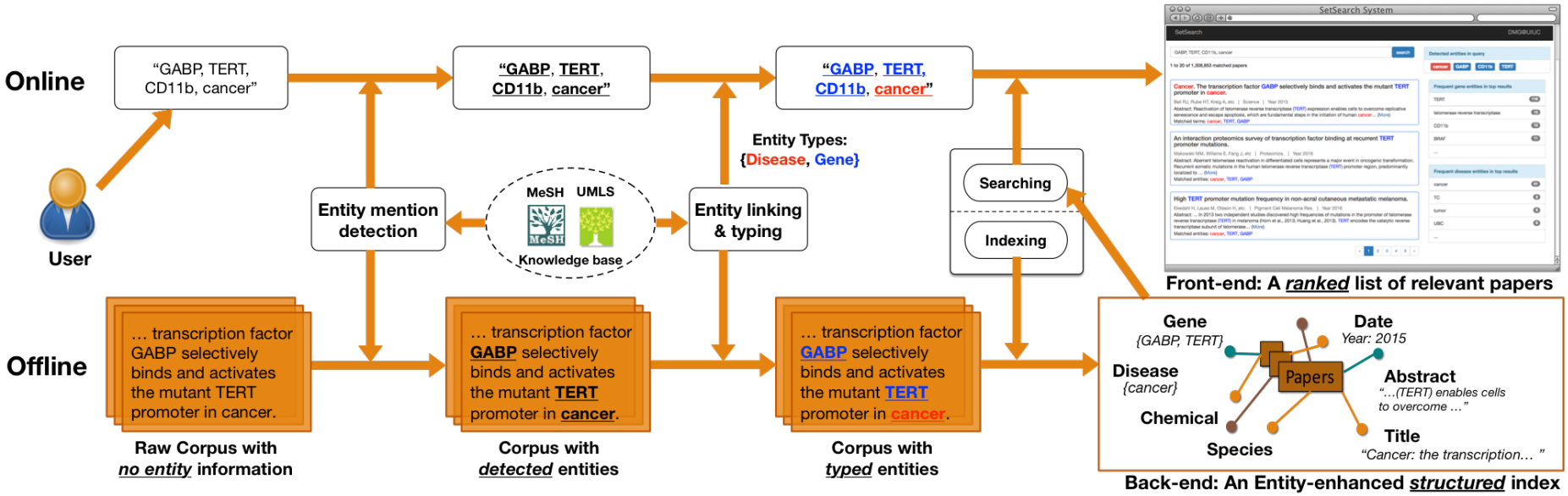
- Time factor: we will favor those more recent papers.
- Field factor: we will favor the entities that present in the title field, compared with those that present in abstract field.

Outline

- Introduction
- Related works
- Proposed method
- System Implementation 
- Results Analysis
- Future works

System Implementation

- System Architecture



System Implementation

- Data (Thanks to Yu Shi)
  - PubMed Articles (26,450,721 papers)
    - Fields: PMID, title, abstract, journal name, author list, publication date
    - Entity: 4 types -- Chemical, Gene, Species, Disease

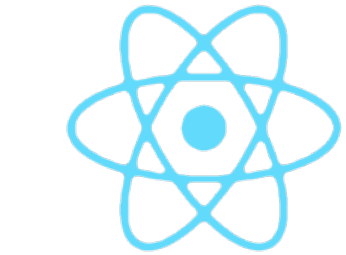
Type	Number of distinct entities	Total occurrence
Chemical	3,279,367	67,678,792
Gene	632,106	30,860,984
Species	79,117	51,730,932
Disease	3,608,247	79,961,123

System Implementation

- Platform
  - Backend search engine
    - Offline Indexing: approximately 6 hours.
    - Online Search: on average < 2s (varies on query)
  - Frontend visualization (Thanks to Jinda Han)
    - Flask + React

elasticsearch

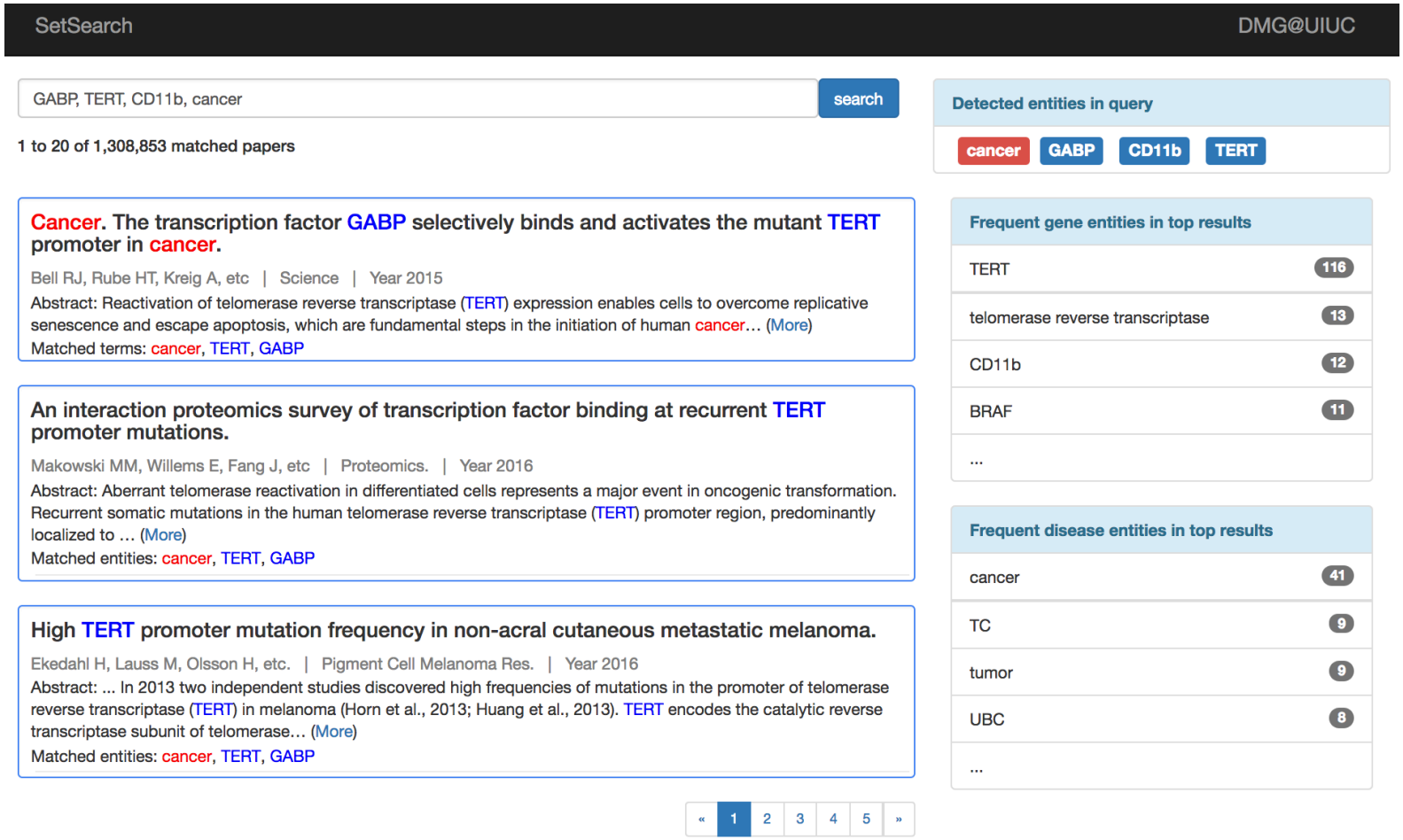
Flask
web development,
one drop at a time



React

System Implementation

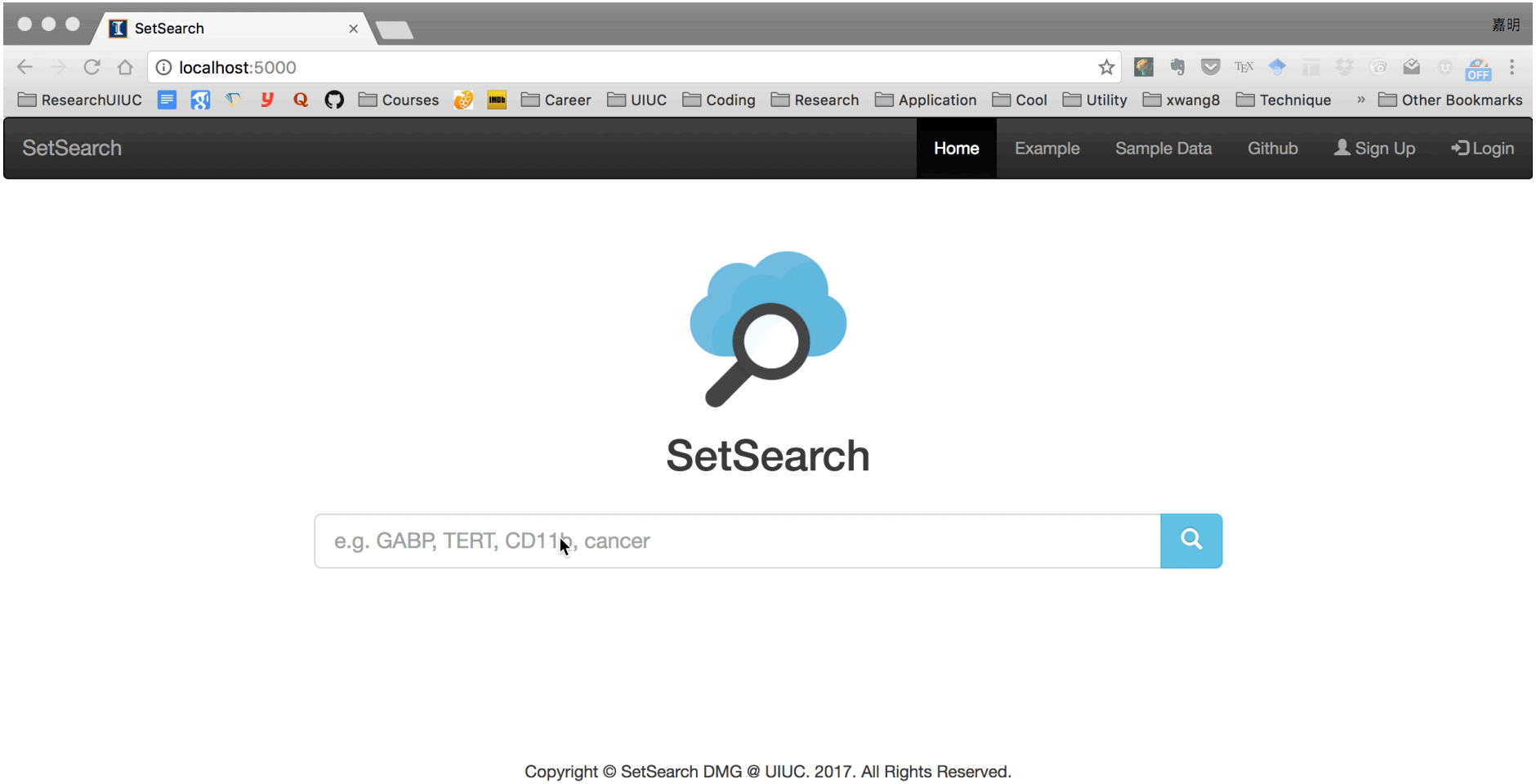
- Interface Prototype



System Implementation

Demo

System Implementation



Outline

- Introduction
- Related works
- Proposed method
- System Implementation
- Results Analysis 
- Future works

Result Analysis

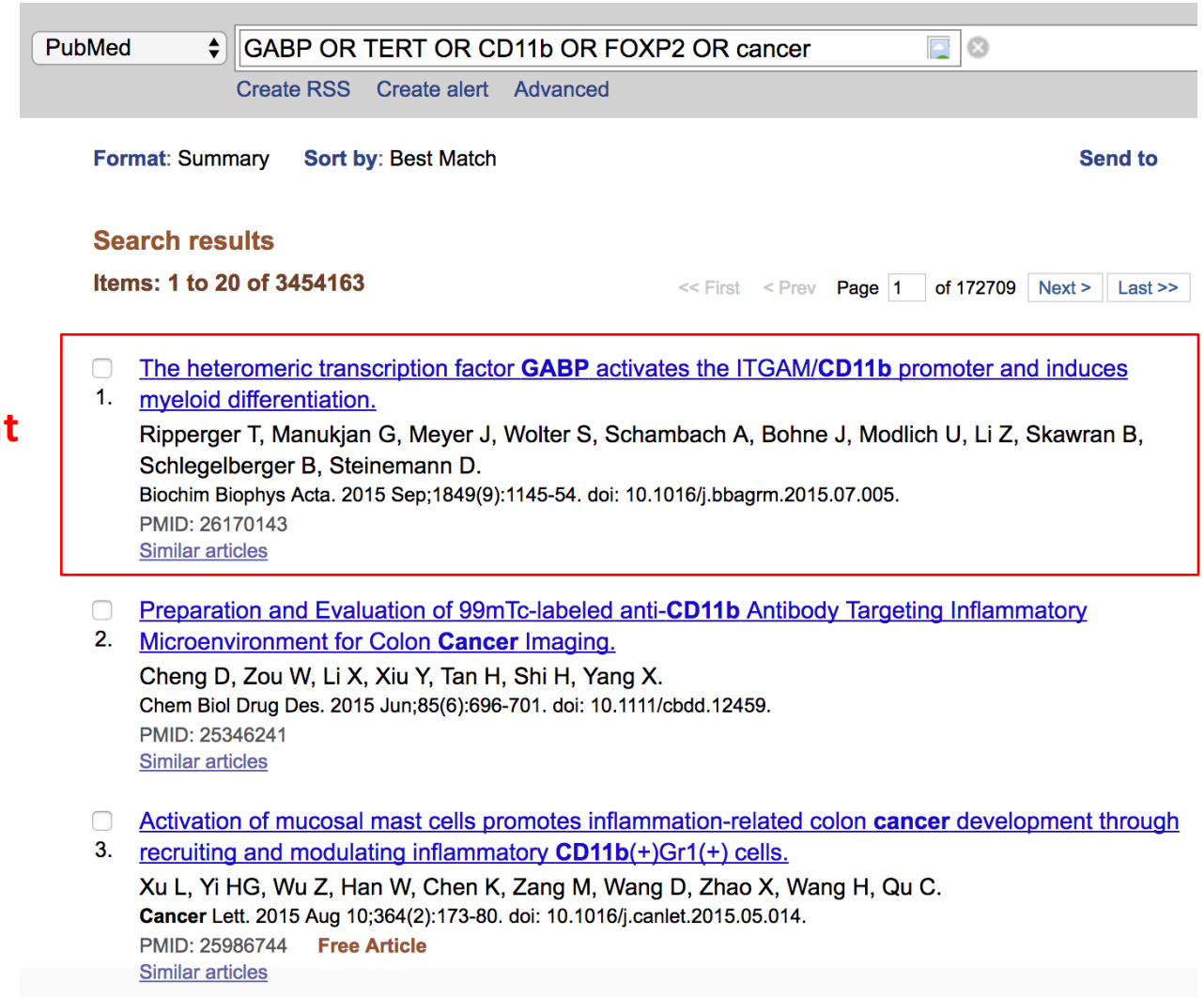
- We use three cases to show the results. (Thanks to Jinfeng Xiao)
  - Query
    - The exact input query of search engine
  - Related question
    - The bio-medical question that users would like to answer with literature search
  - Judging Criterion
    - What kind of papers is considered to be relevant?
    - What order do we prefer among relevant papers?

Case 1

- Query: "*GABP, TERT, CD11b, FOXP2, cancer*"
- Related question: How are the genes *GABP*, *TERT*, *CD11b* and *FOXP2*, as well as their interplay, associated with *cancer*?
- Judging Criterion:
  - A paper is judged as relevant if and only if it covers the relation between *cancer* and at least one gene among *GABP*, *TERT*, *CD11b* and *FOXP2*.
  - Among relevant papers, we prefer to rank higher those covering more *unique* genes.

Performed between
Jan. 12 – 15, 2017

Case 1 - PubMed



Not Relevant (no cancer)

Relevant

Relevant

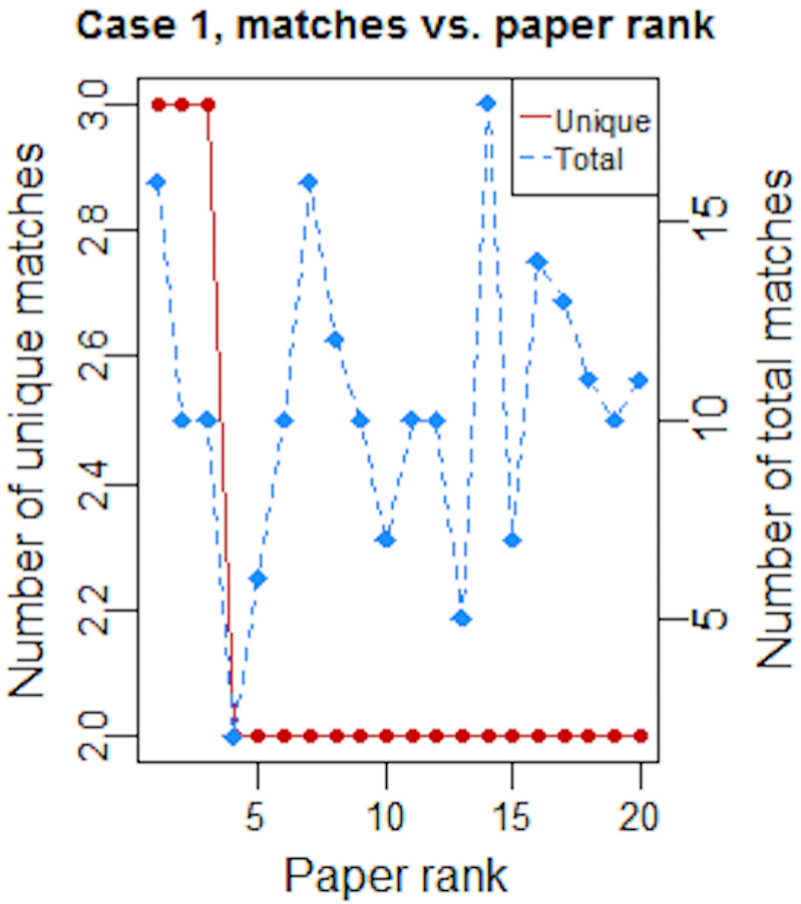
Case 1 - SetSearch

- Paper ranking results (all relevant)

Rank	Title	Date
1	Cancer. The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer.	2015
2	An interaction proteomics survey of transcription factor binding at recurrent TERT promoter mutations.	2016
3	High TERT promoter mutation frequency in non-acral cutaneous metastatic melanoma.	2016
4	TERT promoter mutations in pancreatic endocrine tumours are rare and mainly found in tumours from patients with hereditary syndromes.	2016
5	Silencing FOXP2 in breast cancer cells promotes cancer stem cell traits and metastasis.	2016

Case 1 - SetSearch

- Number of unique terms & total occurrence V.S. paper rank



Case 2

- Query: "*APP, APOE, PSEN1, SORL1, PSEN2, ACE, CLU, BDNF, IL1B, MAPT*" (a set of genes related to Alzheimer's disease [1])
- Related question: What is the underlying context of those 10 genes?
- Judging Criterion:
  - A retrieved paper is relevant if and only if it is discussing at least one of the query genes in the context of Alzheimer's disease.
  - Among relevant papers, we prefer those covering more unique genes.

[1] Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 2016.

Performed between
Jan. 12 – 15, 2017

Case 2 - PubMed

Search results

Items: 1 to 20 of 95197

<< First < Prev Page 1 of 4760 Next > Last >>

Relevant

1. Electroacupuncture at the Baihui acupoint alleviates cognitive impairment and exerts neuroprotective effects by modulating the expression and processing of **brain-derived neurotrophic factor** in **APP**/PS1 transgenic mice.
Lin R, Chen J, Li X, Mao J, Wu Y, Zhuo P, Zhang Y, Liu W, Huang J, Tao J, Chen LD.
Mol Med Rep. 2016 Feb;13(2):1611-7. doi: 10.3892/mmr.2015.4751.
PMID: 26739187 **Free PMC Article**
Similar articles

Not Relevant

2. Synergistic associations of catechol-O-methyltransferase and **brain-derived neurotrophic factor** with executive function in aging are selective and modified by apolipoprotein E.
Sapkota S, Vergote D, Westaway D, Jhamandas J, Dixon RA.
Neurobiol Aging. 2015 Jan;36(1):249-56. doi: 10.1016/j.neurobiolaging.2014.06.020.
PMID: 25107496 **Free PMC Article**
Similar articles

Not Relevant

3. **Brain-Derived Neurotrophic Factor (BDNF)** in Traumatic Brain Injury-Related Mortality: Interrelationships Between Genetics and Acute Systemic and Central Nervous System **BDNF** Profiles.
Failla MD, Conley YP, Wagner AK.
Neurorehabil Neural Repair. 2016 Jan;30(1):83-93. doi: 10.1177/1545968315586465.
PMID: 25979196 **Free PMC Article**
Similar articles

Case 2 - SetSearch

- Paper ranking results (all relevant)

Rank	Title	Date
1	Investigating the role of rare coding variability in Mendelian dementia genes (APP, PSEN1, PSEN2, GRN, MAPT, and PRNP) in late-onset Alzheimer's disease.	2014
2	Screening of Early and Late Onset Alzheimer's Disease Genetic Risk Factors in a Cohort of Dementia Patients from Liguria, Italy	2015
3	Genetics of Alzheimer's disease.	2014
4	The genetics of Alzheimer's disease.	2014
5	The PSEN1, p.E318G variant increases the risk of Alzheimer's disease in APOE- 4 carriers.	2013

All cover the latent entity – Alzheimer's disease

Result Analysis

- Quantitative results for all 3 cases.

Table 1: Precision@20 in three types of queries.

	Google Scholar	PubMed	SetSearch
Case 1	0%	75%	**100%**
Case 2	5%	20%	**100%**
Case 3	5%	15%	**100%**

Outline

- Introduction
- Related works
- Proposed method
- System Implementation
- Results Analysis
- Future works

Future works

- Systematic evaluation of our method
  - More case studies
  - TREC Gene Track datasets
- System Implementation
  - Function Implementations
    - Data updates (incrementally crawl, user upload)
    - Search speed (better server, distributed search)
  - Interface & Visualization (more collaborations)
- More researches on **ontology-guided search**.

Thanks!

Q & A