

Hyperlink Classification via Structured Graph Embedding

*Geon Lee¹, Seonggoo Kang², and Joyce Jiyoung Whang^{*1}*

*¹Sungkyunkwan University (SKKU), ²Naver Corporation, *corresponding author*

The 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019

Contents

1. Outline
2. Datasets and Challenges
3. Knowledge Graph Embedding
4. Hyperlink Classification Model
5. Result
6. Conclusion and Future Work

Outline

Three types of hyperlinks:

Navigation Link

Links related to navigation within a site, such as site path or site recommendation.

Suggestion Link

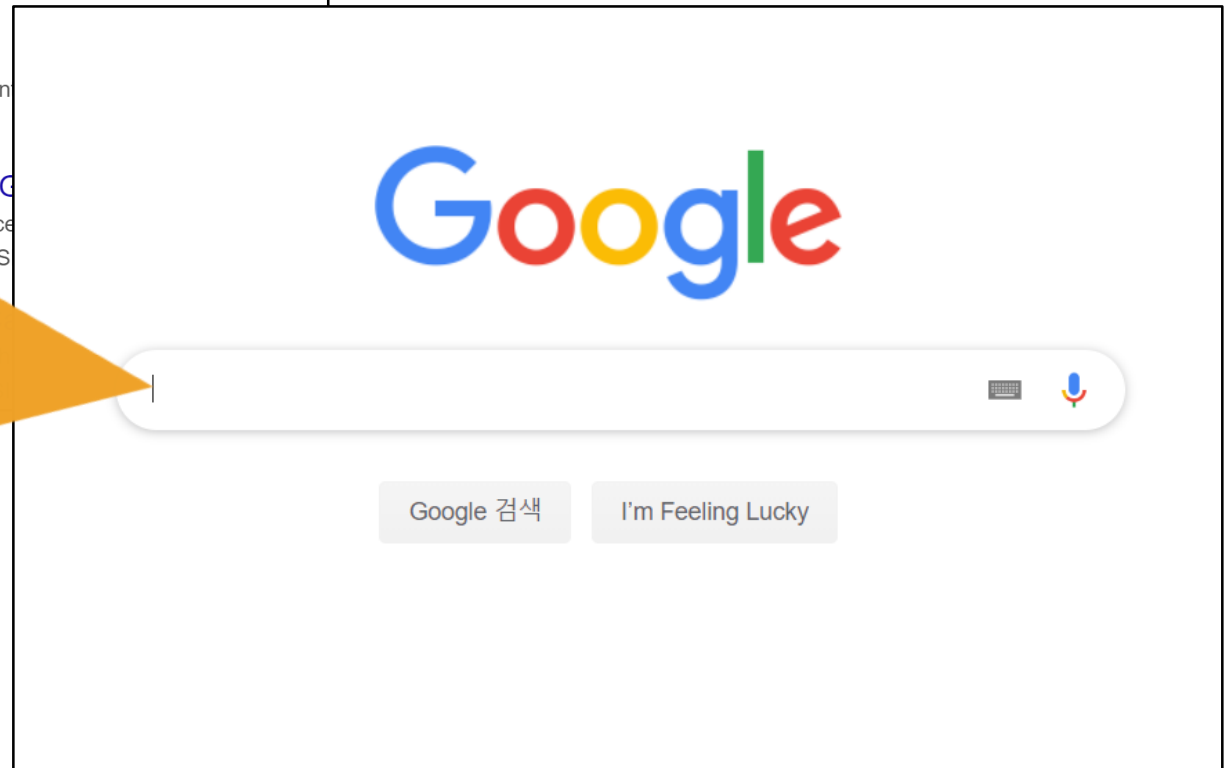
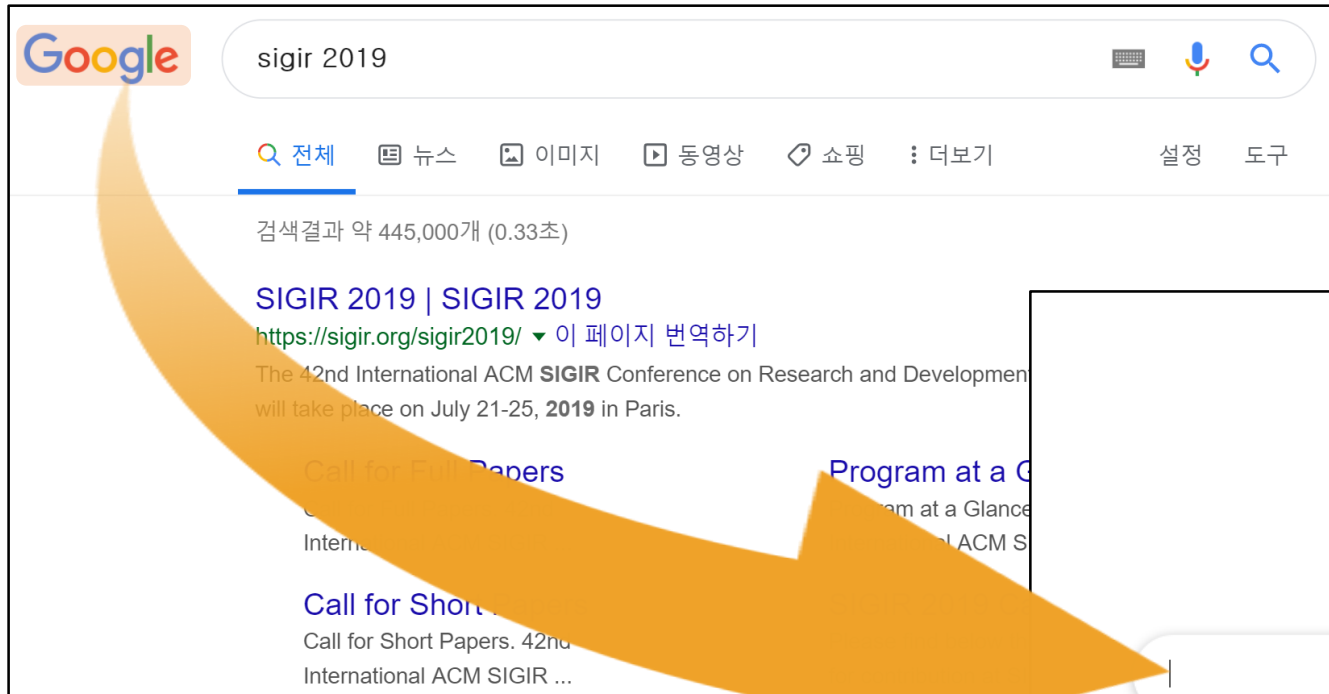
Links recommended directly by the user using the site.

Action Link

Links that require other user actions.

Outline

Navigation Link



Outline

Suggestion Link

Stack Overflow

Search...

Home

PUBLIC

Stack Overflow

Tags

Users

Jobs

Teams

Q&A for work

Learn More

show 1 more comment

3 Answers

active oldest votes

4

If you need the inverse, you should use `inv`.

The inverse is calculated via LU decomposition, whereas the backslash operator `\` calculates the solution to your linear system using different methods depending on the properties of your matrix `A` (see <https://scicomp.stackexchange.com/a/1004>), which can yield less accurate results for the inverse.

It should be noted that if you want to solve a linear system, the calculation is likely going to be much faster and more accurate using `\`. The MATLAB documentation of `inv` is basically one big warning not to use `inv` to solve linear systems.

share improve this answer

edited Apr 13 '17 at 12:53

answered Jan 27 '16 at 10:26

Community 1 • 1

dasdingonesin 1,268 • 1 • 8 • 15

add a comment

More jobs means more choice

Let's disregard performance (speed) and best practice for a bit.

7

`eps(n)` is a command that returns the distance to the next larger double precision number from `n` in MATLAB. So, `eps(1) = 2.2204e-16` means that the first number after `1` is `1 + 2.2204e-16`. Similarly, `eps(3000) = 4.5475e-13`. Now, let's look at the precision of your calculations:

```
n = 100;
A = rand(n);
inv_A_1 = inv(A);
inv_A_2 = A \ eye(n);

max(max(abs(inv_A_1-inv_A_2)))
ans =
    1.6431e-14
```

Stack Exchange

Search on Computational Science...

Home

Questions

Tags

Users

Unanswered

37

In Matlab, the `\` command invokes an algorithm which depends upon the structure of the matrix `A` and includes checks (small overhead) on properties of `A`.

1. If `A` is sparse and banded, employ a banded solver.
2. If `A` is an upper or lower triangular matrix, employ a backward substitution algorithm.
3. If `A` is symmetric and has real positive diagonal elements, attempt a Cholesky factorization. If `A` is sparse, employ reordering first to minimize fill-in.
4. If none of criteria above is fulfilled, do a general triangular factorization using Gaussian elimination with partial pivoting.
5. If `A` is sparse, then employ the UMFPACK library.
6. If `A` is not square, employ algorithms based on QR factorization for undetermined systems.

To reduce overhead it is possible to use the `linsolve` command in Matlab and select a suitable solver among these options yourself.

share cite improve this answer

answered Jan 25 '12 at 20:06

Allan P. Engsig-Karup 2,091 • 12 • 28

Assuming I am dealing with a 10000x10000 unstructured dense matrix with all elements nonzero (High level of density), what would be my best bet? I want to isolate that 1 algorithm which works for dense matrices. Is it LU, QR or Gaussian Elimination? – Inquest Jan 25 '12 at 20:14

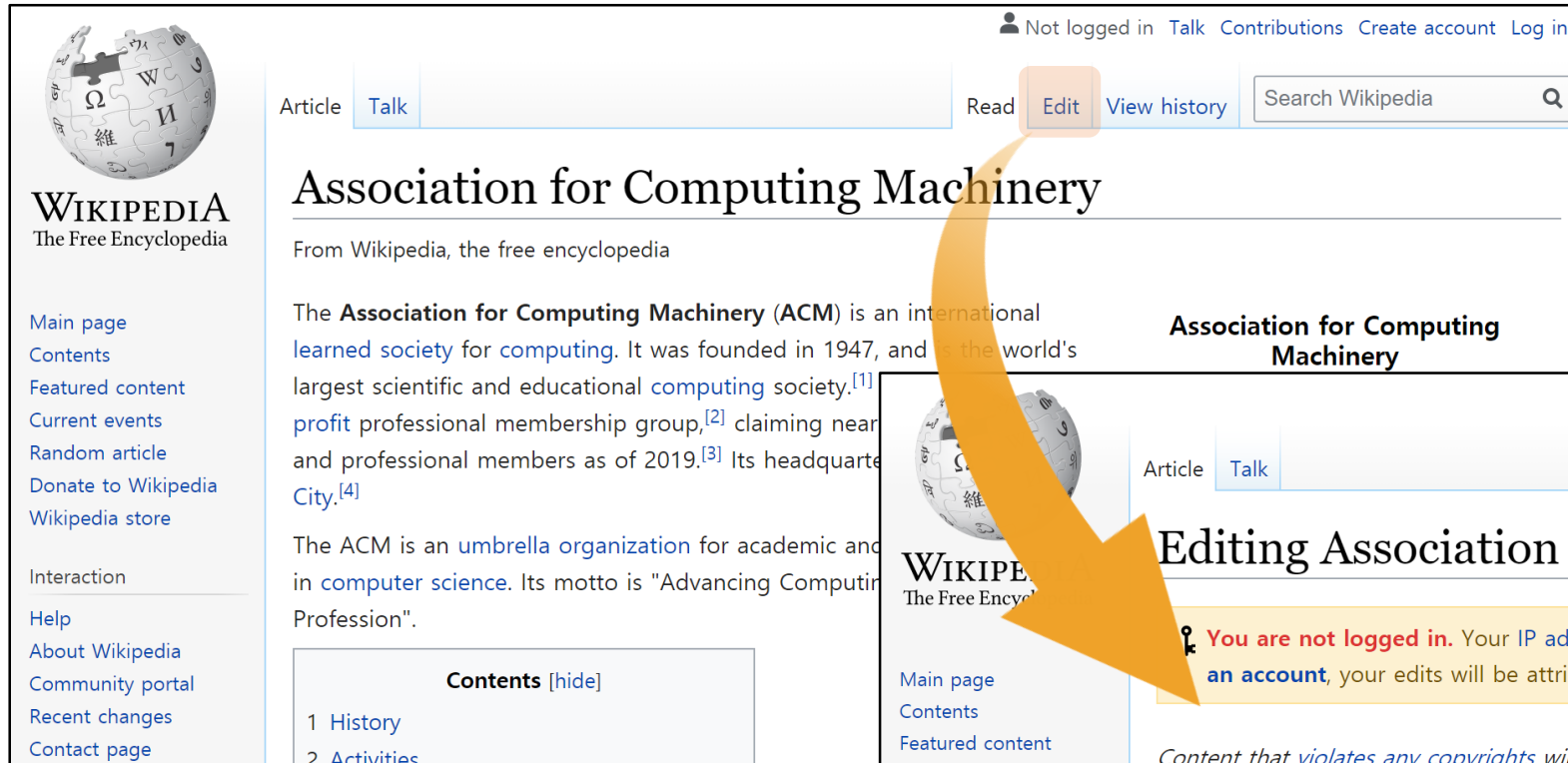
- 1 Sounds like a Step 4 where Gaussian Elimination is invoked which corresponds to the most general case where no structure of `A` can be exploited to boost performance. So, basically this is a LU factorization and subsequent one forward followed by a backward substitution step. – Allan P. Engsig-Karup Jan 25 '12 at 20:18

Thanks! I think that gives me a direction to think. Currently, Gaussian Elimination is the best we have for solving such unstructured problems, is that correct? – Inquest Jan 25 '12 at 20:25

add a comment

Outline

Action Link



This screenshot shows the Wikipedia article page for "Association for Computing Machinery". The page is viewed in a desktop browser. The top navigation bar includes links for "Not logged in", "Talk", "Contributions", "Create account", and "Log in". Below this, the article title "Association for Computing Machinery" is displayed. The "Edit" button is highlighted with an orange box. The article text describes the ACM as an international learned society for computing, founded in 1947, and the world's largest scientific and educational computing society. It mentions that it is a profit professional membership group, claiming nearly 100,000 members as of 2019. Its headquarters are in New York City. The ACM is an umbrella organization for academic and professional computing in computer science. Its motto is "Advancing Computing for the Benefit of the Profession".

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Article Talk

Read Edit View history

Search Wikipedia

Association for Computing Machinery

From Wikipedia, the free encyclopedia

The **Association for Computing Machinery (ACM)** is an international learned society for computing. It was founded in 1947, and is the world's largest scientific and educational computing society.^[1] It is a profit professional membership group,^[2] claiming nearly 100,000 members as of 2019.^[3] Its headquarters are in New York City.^[4]

The ACM is an umbrella organization for academic and professional computing in computer science. Its motto is "Advancing Computing for the Benefit of the Profession".

Contents [hide]

- History
- Activities



This screenshot shows the Wikipedia editing page for "Association for Computing Machinery". The page is viewed in a desktop browser. The top navigation bar includes links for "Not logged in", "Talk", "Contributions", "Create account", and "Log in". Below this, the article title "Association for Computing Machinery" is displayed. The "Edit source" button is highlighted with an orange box. A yellow warning box states: "You are not logged in. Your IP address will be publicly visible if you make any edits. If you log in or create an account, your edits will be attributed to a user name, among other benefits." The article text describes the ACM as an international learned society for computing, founded in 1947, and the world's largest scientific and educational computing society. It mentions that it is a profit professional membership group, claiming nearly 100,000 members as of 2019. Its headquarters are in New York City. The ACM is an umbrella organization for academic and professional computing in computer science. Its motto is "Advancing Computing for the Benefit of the Profession".

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Article Talk

Read Edit source View history

Search Wikipedia

Editing Association for Computing Machinery

You are not logged in. Your IP address will be publicly visible if you make any edits. If you **log in** or **create an account**, your edits will be attributed to a user name, among other benefits.

Content that violates any copyrights will be deleted. Encyclopedic content must be verifiable. Work submitted to Wikipedia can be edited, used, and redistributed—by anyone—subject to certain terms and conditions.

B I      **Advanced** **Special characters** **Help** **Cite**

```
{{Use mdy dates|date=June 2012}}
{{Infobox organization
| name       = Association for Computing Machinery
| image      = Association for Computing Machinery (ACM) logo.svg
| image_border = 
| size       = 100px
| alt        = "Logo" is blue circle with "ACM" in white, surrounded by blue diagonal lines
```

Datasets and Challenges

Conduct a biased random walk from the seed page. (By NAVER)

The screenshot shows a Stack Overflow page with the title "Why is it faster to process a sorted array than an unsorted array?". The page has a left sidebar with navigation links (Home, PUBLIC, Stack Overflow, Tags, Users, Jobs, Teams) and a search bar. The main content area displays a question with a C++ code snippet. The code generates an array of 32768 random numbers and compares the execution time of a loop that sorts the array versus a loop that does not. The code is as follows:

```
#include <algorithm>
#include <ctime>
#include <iostream>

int main()
{
    // Generate data
    const unsigned arraySize = 32768;
    int data[arraySize];

    for (unsigned c = 0; c < arraySize; ++c)
        data[c] = std::rand() % 256;

    // !!! With this, the next loop runs faster
    std::sort(data, data + arraySize);

    // Test
    clock_t start = clock();
    long long sum = 0;

    for (unsigned i = 0; i < 100000; ++i)
    {
        // Primary loop
        for (unsigned c = 0; c < arraySize; ++c)
        {
            if (data[c] >= 128)
                sum += data[c];
        }
    }

    double elapsedTime = static_cast<double>(clock() - start) / CLOCKS_PER_SEC;

    std::cout << elapsedTime << std::endl;
    std::cout << "sum = " << sum << std::endl;
}
```

Below the code, there are two bullet points:

- Without `std::sort(data, data + arraySize);`, the code runs in 11.54 seconds.
- With the sorted data, the code runs in 1.93 seconds.

The question text states: "Initially, I thought this might be just a language or compiler anomaly. So I tried it in Java."

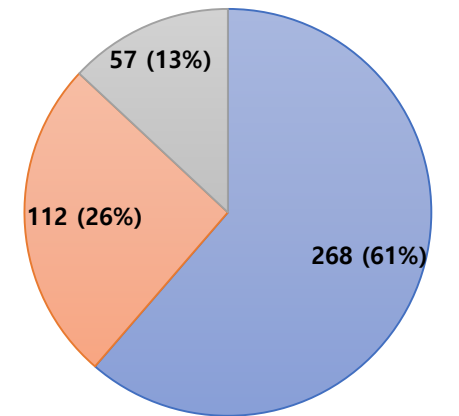
On the right side of the page, there is a "Looking for a job?" section with several job listings, including "Cloud- Python Developer", "Back-end(Python) Developer", "Front End Developer", and "Senior iOS Engineer".

Seed Page

Datasets and Challenges

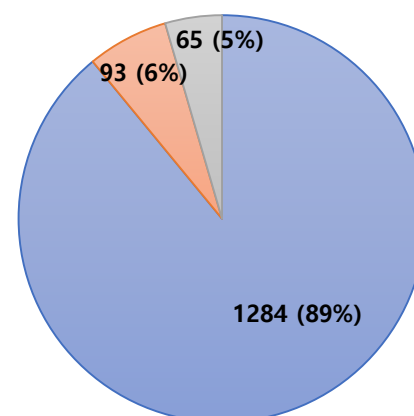
Datasets

	graph_437	graph_1442	Graph_10000
No. of Hyperlinks	437	1442	10000
No. of Pages	404	332	2202
No. of Domain	18	100	120
No. of Host	-	22	25



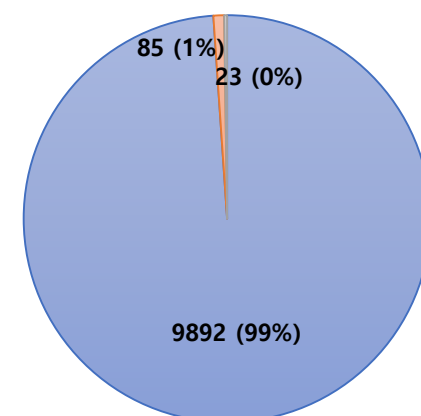
Navigation Suggestion Action

graph_437



Navigation Suggestion Action

graph_1442

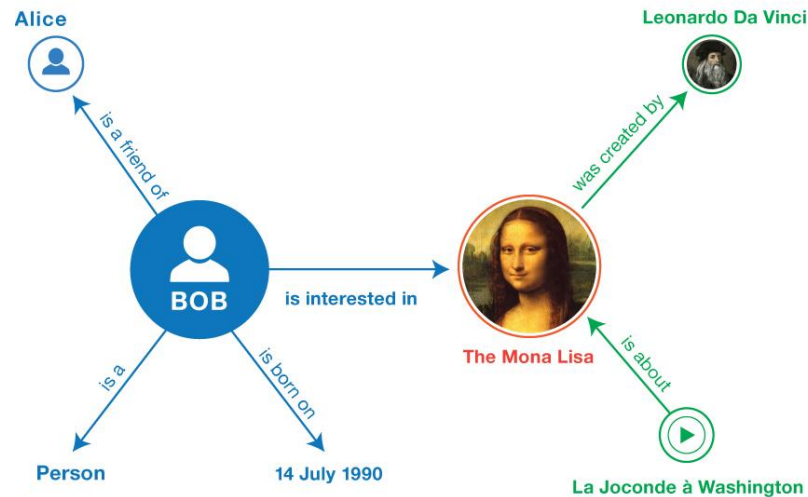


Navigation Suggestion Action

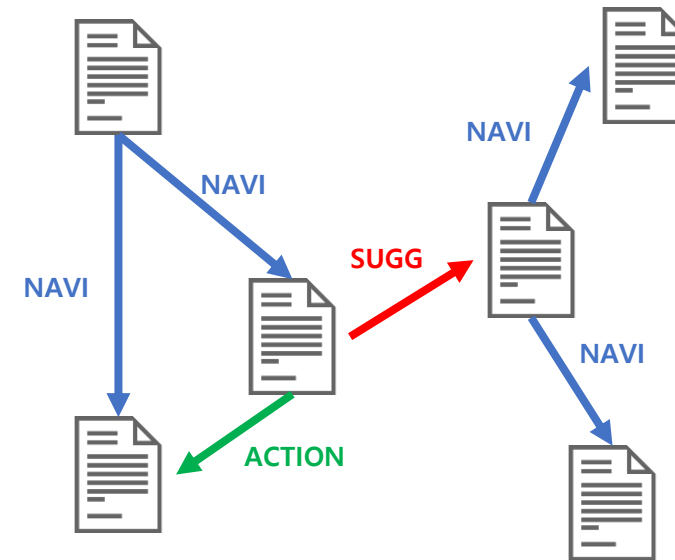
graph_10000

Knowledge Graph Embedding

Can the **knowledge graph embedding** method be applied to **hyperlink embedding**?



Knowledge Graph



Web Graph

Knowledge Graph Embedding

Can the **knowledge graph embedding** method be applied to **hyperlink embedding**?

(Bob, **is_interested_in**, The_Mona_Lisa)

(Bob, **_is_a_friend_of**, Alice)

(Barack Obama, **_place_of_birth**, Hawaii)

(Albert Einstein, **_follows_diet**, Veganism)

(San Francisco, **_contains**, Telegraph Hill)



(www.naver.com, **NAVI**, www.news.naver.com)

(www.google.com/larry_page, **SUGG**, www.wikipedia.org/Larry_Page)

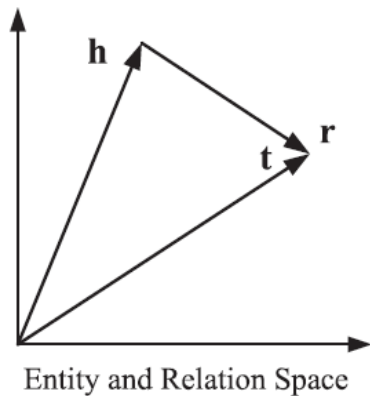
(www.wikipedia.org/Larry_Page, **ACTION**, www.wikipedia.org/Larry_Page/edit)

Knowledge Graph Embedding

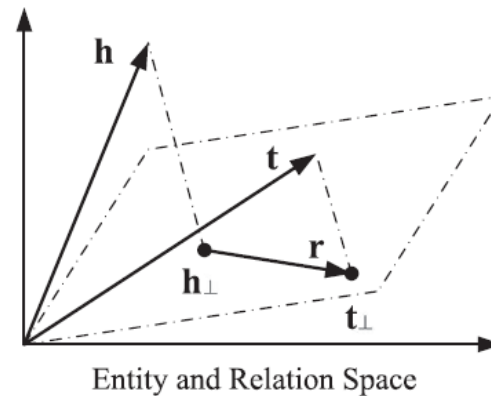
TransE : Translating Embeddings for Modeling Multi-relational Data (NIPS 2013)

TransH : Knowledge Graph Embedding by Translating on Hyperplanes (AAAI 2014)

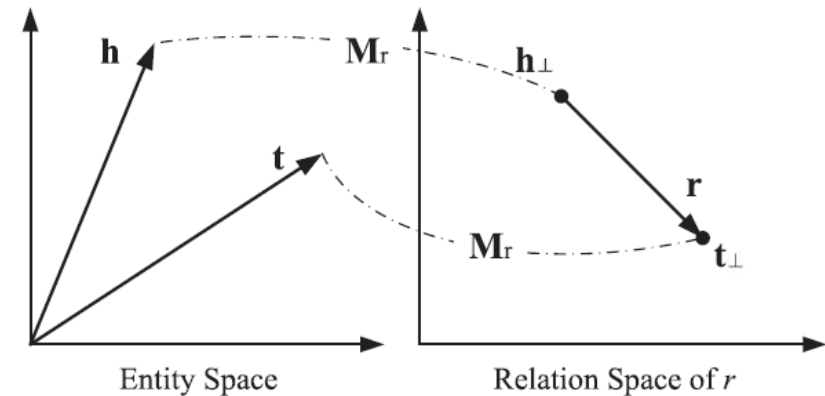
TransR : Learning Entity and Relation Embeddings for Knowledge Graph Completion (AAAI 2015)



(a) TransE.



(b) TransH.



(c) TransR.

Knowledge Graph Embedding

TransE, **TransH**, and **TransR** methods minimize the following loss function:

$$L = \sum_{(h,r,t) \in S} \sum_{(h',r,t') \in S'} [f(h,r,t) + \gamma - f(h',r,t')]_+$$

where $[x]_+ \equiv \max(0, x)$ and γ is the margin.

S is a set of golden triplets and S' is a set of corrupted triplets.

The method for calculating $f(h,r,t)$ depends on the model.

$$\textbf{TransE} : f(h,r,t) = \|h + r - t\|_2^2$$

$$\textbf{TransH} : f(h,r,t) = \|h_{\perp} + r - t_{\perp}\|_2^2$$

$$\textbf{TransR} : f(h,r,t) = \|h_r + r - t_r\|_2^2$$

e_{\perp} and e_r represent projected entity e on relation-specific hyperplane and space, respectively.

Hyperlink Classification Model

Given two web pages P_u and P_v , and set of hyperlink types R , estimate the type of hyperlink between P_u and P_v as follow.

$$r^* = \underset{r \in R}{\mathbf{argmin}} f(P_u, r, P_v)$$

where $f(h, r, t)$ is the loss of triplet (h, r, t) .

r^* is the predicted relation.

Hyperlink Classification Model

Negative Sampling

TransE / TransH / TransR → **only changes entities** when making negative samples

$$(h, r, t) \rightarrow (h', r, t) \text{ or } (h, r, t')$$

Introduce parameter α ($0 \leq \alpha \leq 1$).

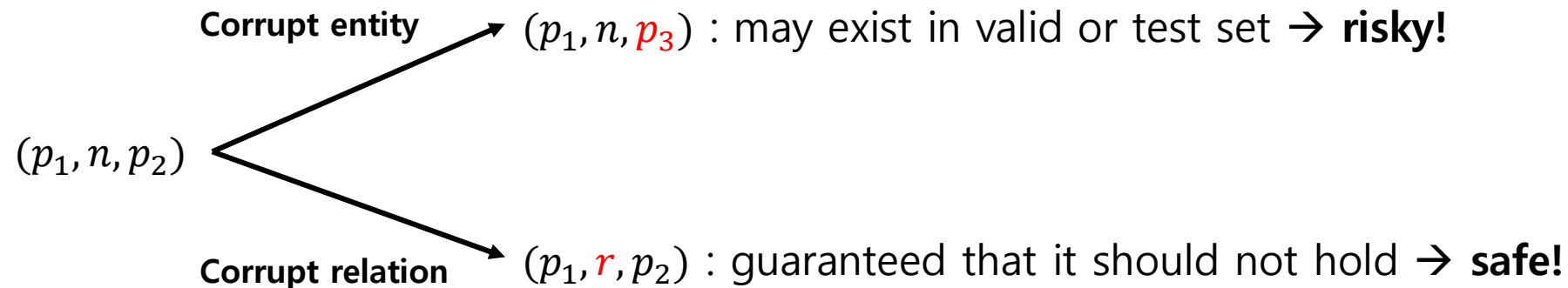
- Replace the **entity** with the probability α . $(h, r, t) \rightarrow (h', r, t) \text{ or } (h, r, t')$
- Replace the **relation** with the probability $1 - \alpha$. $(h, r, t) \rightarrow (h, r', t)$

Hyperlink Classification Model

Negative Sampling

False Negative

When we corrupt entities, there is a chance that it is not a *corrupted* one but just an *unobserved* one in the training set.



Result

The average F1 scores(%) of our model with different α and the original TransE, TransH, and TransR.

		TransE	TransH	TransR
web_437	Our model, $\alpha = 0.3$	34.29	60.25	57.99
	Our model, $\alpha = 0.5$	34.39	58.87	57.32
	Our model, $\alpha = 0.7$	33.88	58.91	59.83
	The original model	36.22	54.04	53.22
web_1442	Our model, $\alpha = 0.3$	23.39	53.42	50.04
	Our model, $\alpha = 0.5$	24.86	55.16	46.18
	Our model, $\alpha = 0.7$	21.18	52.70	45.12
	The original model	20.05	29.94	10.35
web_10000	Our model, $\alpha = 0.3$	20.68	76.00	53.86
	Our model, $\alpha = 0.5$	17.98	74.64	46.99
	Our model, $\alpha = 0.7$	19.50	72.94	44.11
	The original model	15.31	25.35	2.08

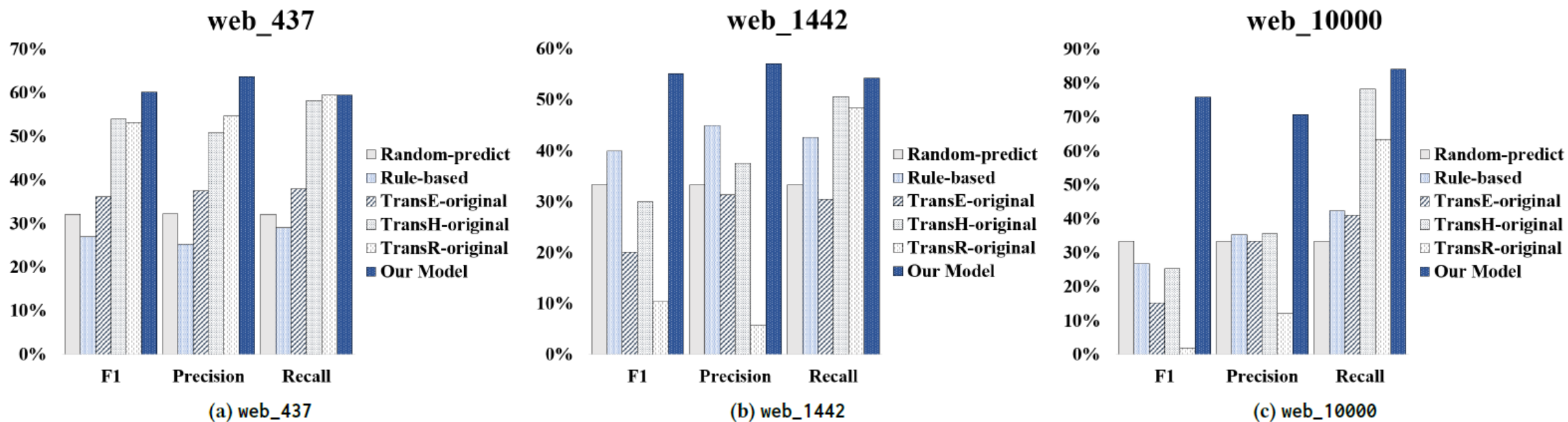
Result

F1 Scores(%) of each relation and the average F1 scores.

		<i>navigation</i>	<i>suggestion</i>	<i>action</i>	Average
web_437	Random-predict	59.75	25.81	11.07	32.21
	Rule-based	60.20	20.96	0.00	27.05
	TransE-original	55.78	31.96	20.93	36.22
	TransH-original	70.80	52.75	38.56	54.04
	TransR-original	67.87	52.86	38.94	53.22
	Our Model	77.04	57.05	46.64	60.25
web_1442	Random-predict	89.13	5.18	5.65	33.32
	Rule-based	72.98	10.20	36.67	39.95
	TransE-original	42.54	8.57	9.05	20.05
	TransH-original	54.80	13.57	21.45	29.94
	TransR-original	0.00	12.97	18.09	10.35
	Our Model	93.48	22.88	49.12	55.16
web_10000	Random-predict	98.91	1.60	0.00	33.50
	Rule-based	68.81	1.74	9.92	26.82
	TransE-original	43.25	2.06	0.61	15.31
	TransH-original	63.01	12.02	1.03	25.35
	TransR-original	0.00	5.61	0.61	2.08
	Our Model	99.66	83.22	45.12	76.00

Result

The average F1, precision, recall on the three web graphs.



- **Random-predict** indicates the performance of random prediction while preserving the number of hyperlinks in each class.
- We use the result of TransH with $\alpha = 0.3$, $\alpha = 0.5$, and $\alpha = 0.3$ for web_437, web_1442, and web_10000, respectively.

Result

Performance on the original web graphs and randomly shuffled graphs where the relation labels are randomly assigned while preserving the number of each relation type.

		<i>navigation</i>			<i>suggestion</i>			<i>action</i>		
		F1 (%)	Pre. (%)	Rec. (%)	F1 (%)	Pre. (%)	Rec. (%)	F1 (%)	Pre. (%)	Rec. (%)
web_437	Original Graph	77.04	78.82	75.37	57.05	50.43	65.77	46.64	62.00	37.43
	Randomly Shuffled Graph	58.60	60.51	56.88	25.36	24.39	26.59	13.79	13.26	14.42
web_1442	Original Graph	93.48	92.22	94.78	22.88	30.66	18.28	49.12	48.52	49.74
	Randomly Shuffled Graph	86.08	88.94	83.41	6.19	5.28	7.53	5.68	4.58	7.52
web_10000	Original Graph	99.66	99.82	99.50	83.22	77.84	89.41	45.12	34.91	63.77
	Randomly Shuffled Graph	98.43	98.94	97.92	1.28	0.99	1.83	0.61	0.38	1.45

The real-world web graphs have characterized structures, which enables the knowledge graph embedding techniques to reasonably work well on predicting the relation labels.

Conclusion and Future Work

Conclusion

- The knowledge graph embedding techniques can be efficiently used for hyperlink embedding.
- Performance improvement can be observed by modifying the negative sampling method.

Future Work

- Utilize information from web pages or hyperlinks (such as anchor text) in hyperlink embeddings.

Thank You