Hyperlink Classification via Structured Graph Embedding

$Geon\ Lee^{1}$, $Seonggoo\ Kang^{2}$, and $Joyce\ Jiyoung\ Whang^{*1}$

$^{1}Sungkyunkwan\ University\ (SKKU)$, $^{2}Naver\ Corporation$, *corresponding author

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

Contents

Outline

Three types of hyperlinks:

Navigation Link

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

Suggestion Link

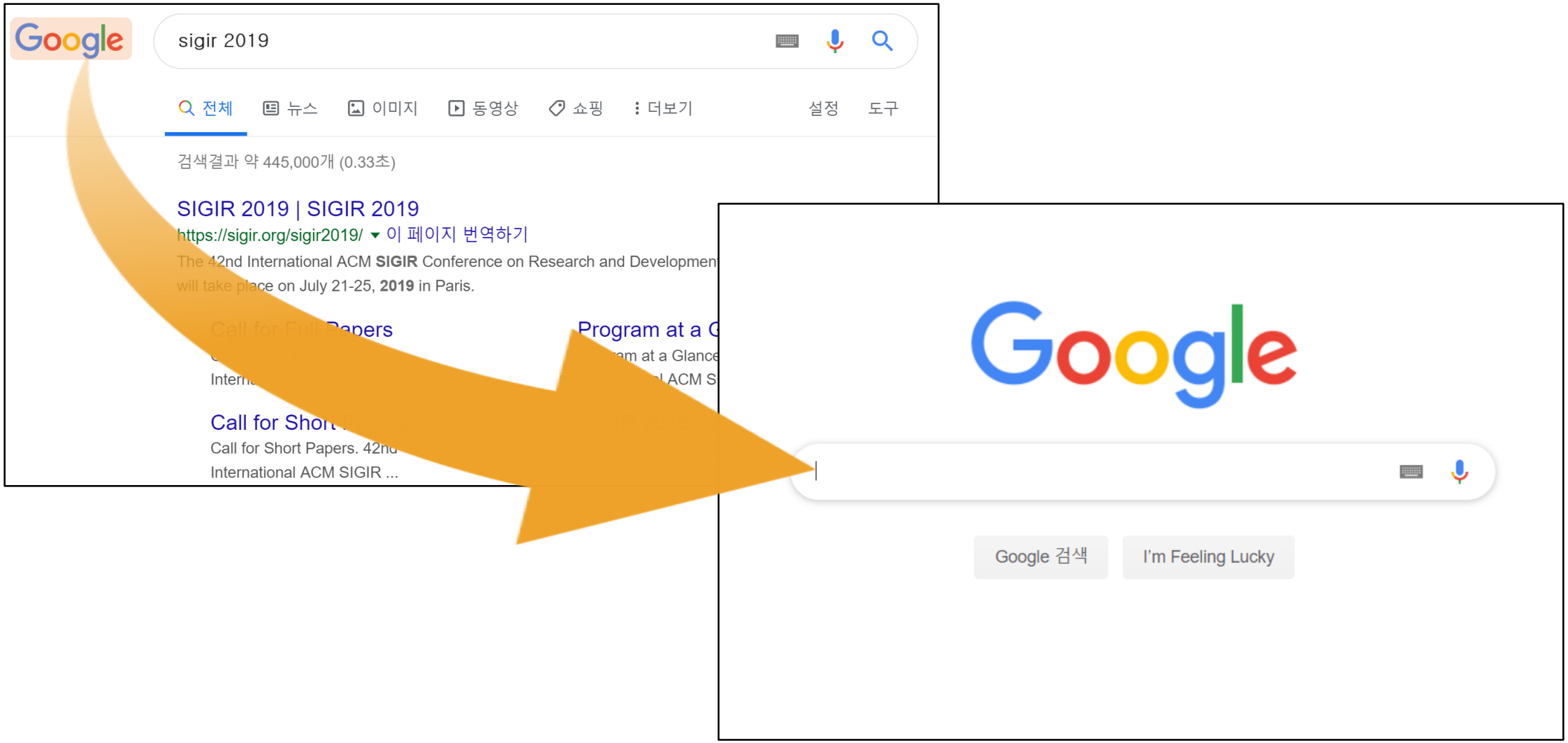
Links recommended directly by the user using the site.

Action Link

Links that require other user actions.

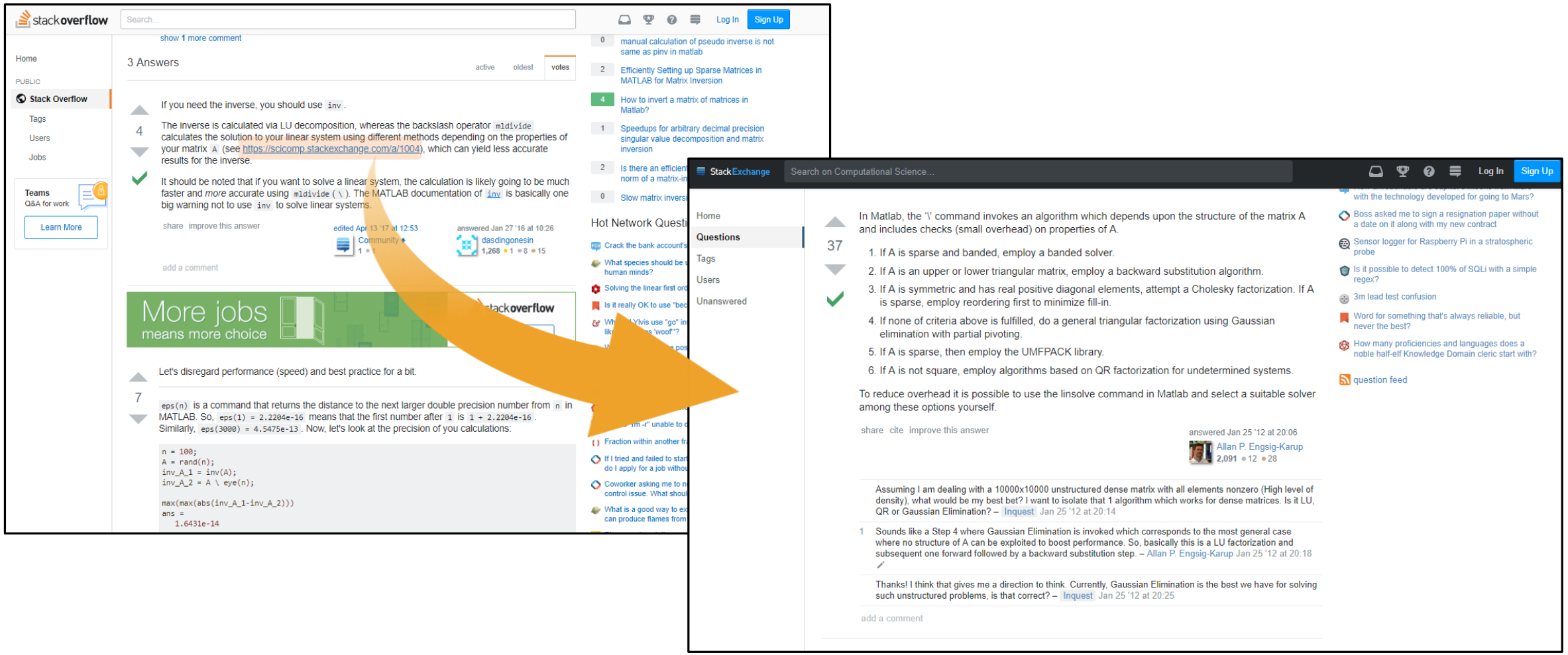
Outline

Navigation Link



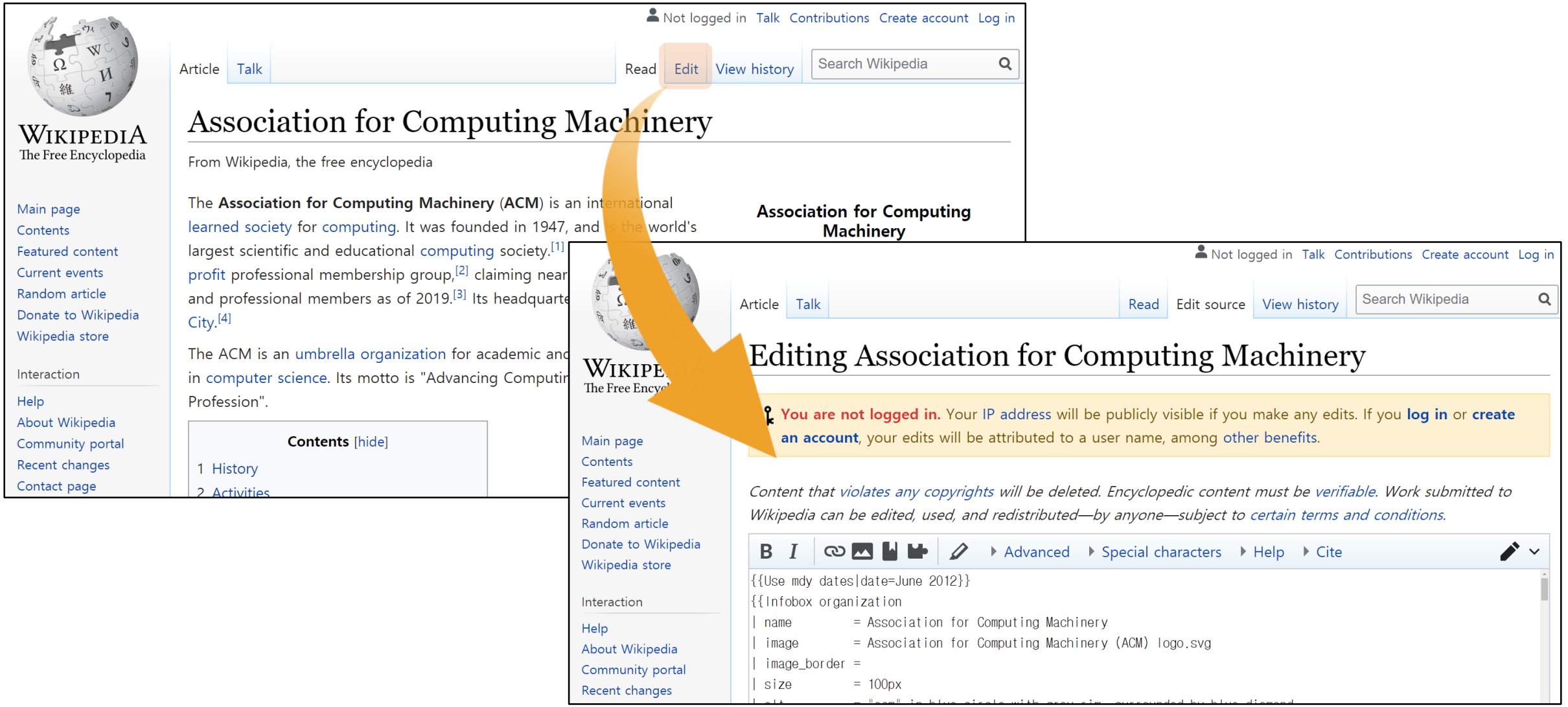
Outline

Suggestion Link



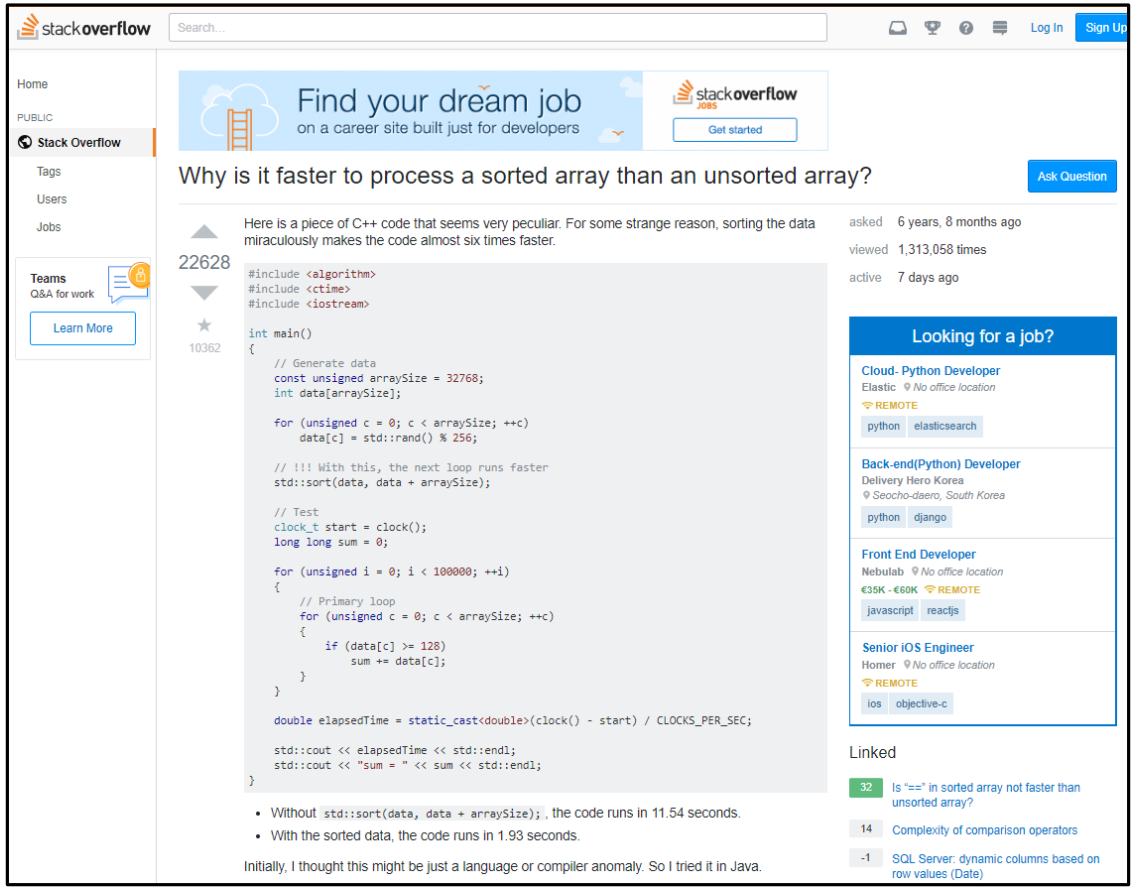
Outline

Action Link



Datasets and Challenges

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

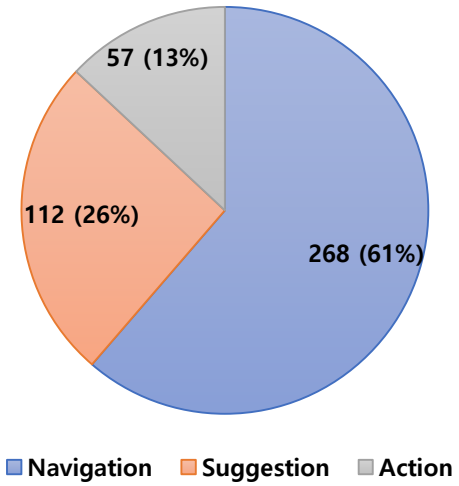


Seed Page

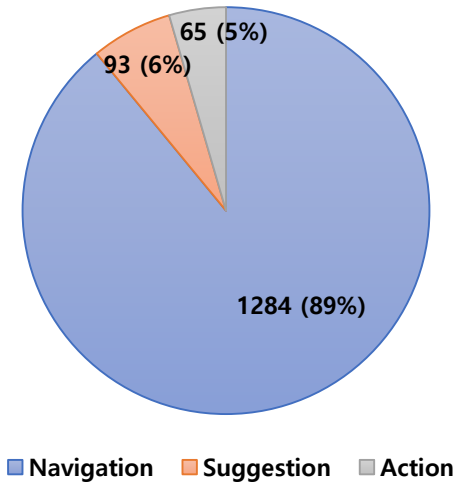
Datasets and Challenges

Datasets

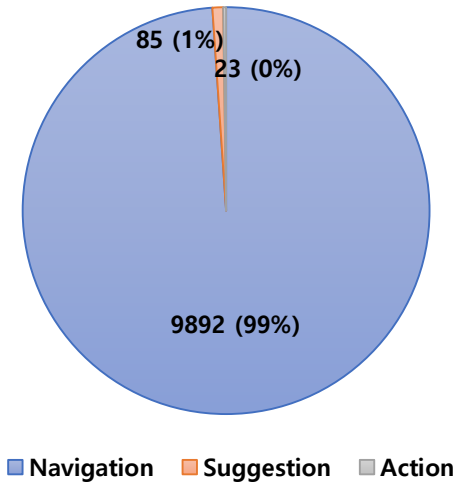
	graph_437	graph_1442	Graph_10000
No. of Hyperlinks	437	1442	10000
No. of Pages	404	332	2202
No. of Domain	18	100	120
No. of Host	-	22	25



graph_437



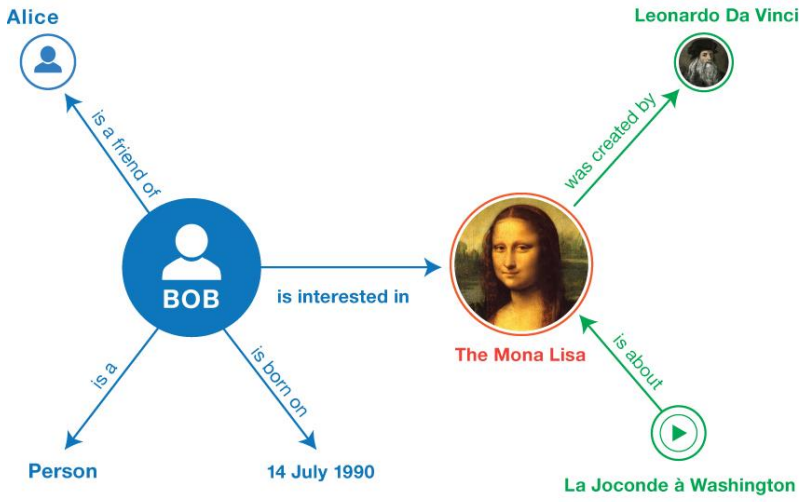
graph_1442



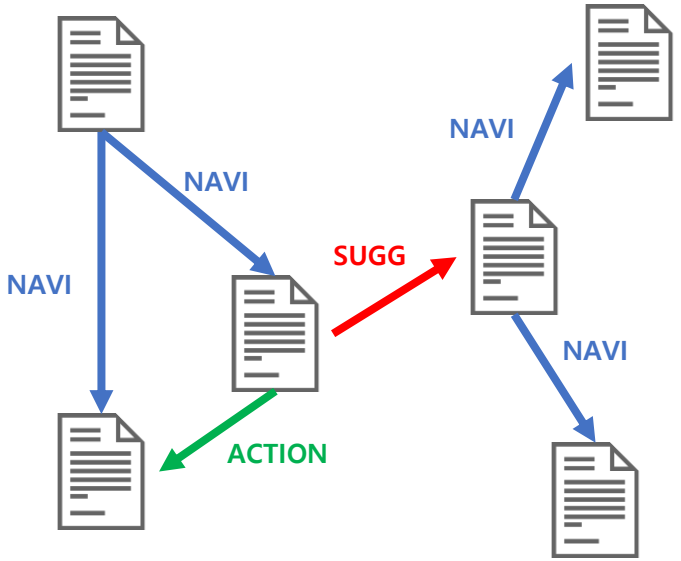
graph_10000

Knowledge Graph Embedding

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



Knowledge Graph



Web Graph

Knowledge graph image from https://medium.com/comet-ml/using-fasttext-and-comet-ml-to-classify-relationships-in-knowledge-graphs-e73d27b40d67

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)

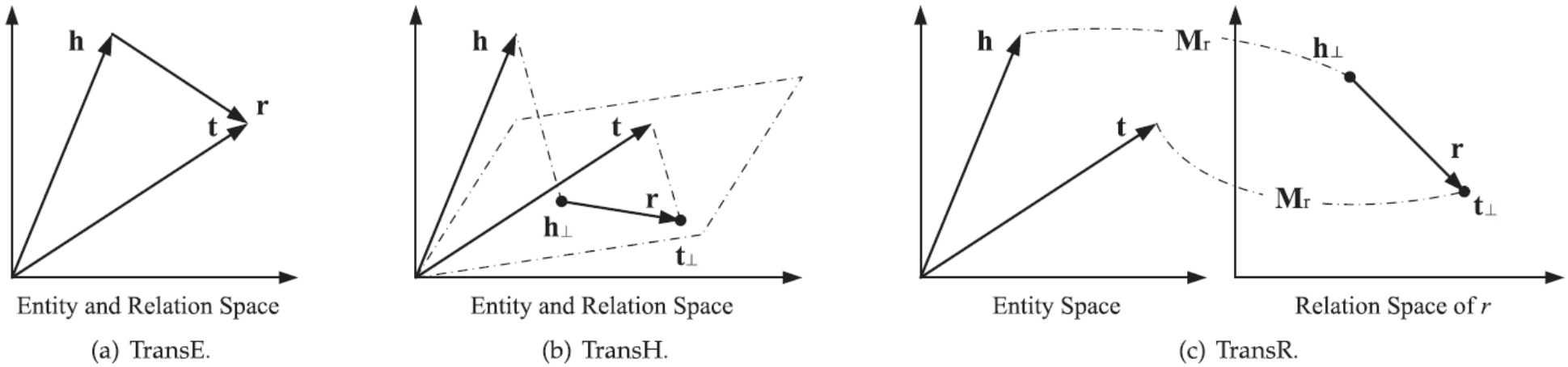


Image from "Knowledge graph embedding: A survey of approaches and applications.", TKDE 2017.

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.

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

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

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.

$$r^* = \boldsymbol{argmin}_{r \in \mathrm{R}} f(\boldsymbol{P_u}, r, \boldsymbol{P_v})$$

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 $(0 \leq \alpha \leq 1)$.

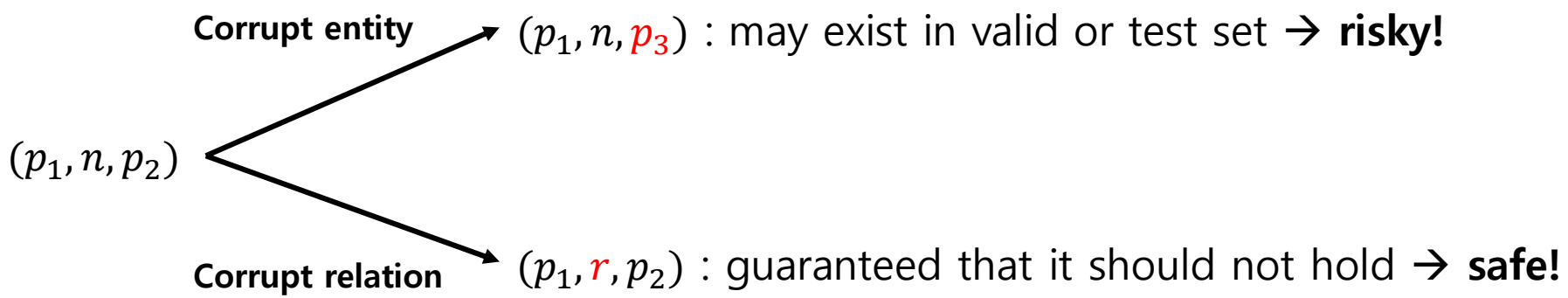
- Replace the **entity** with the probability α. $(h, r, t) \rightarrow (h', r, t)$ or (h, r, t')
- Replace the **relation** with the probability $1 - \alpha$. $(h, r, t) \rightarrow (h, r', t)$

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

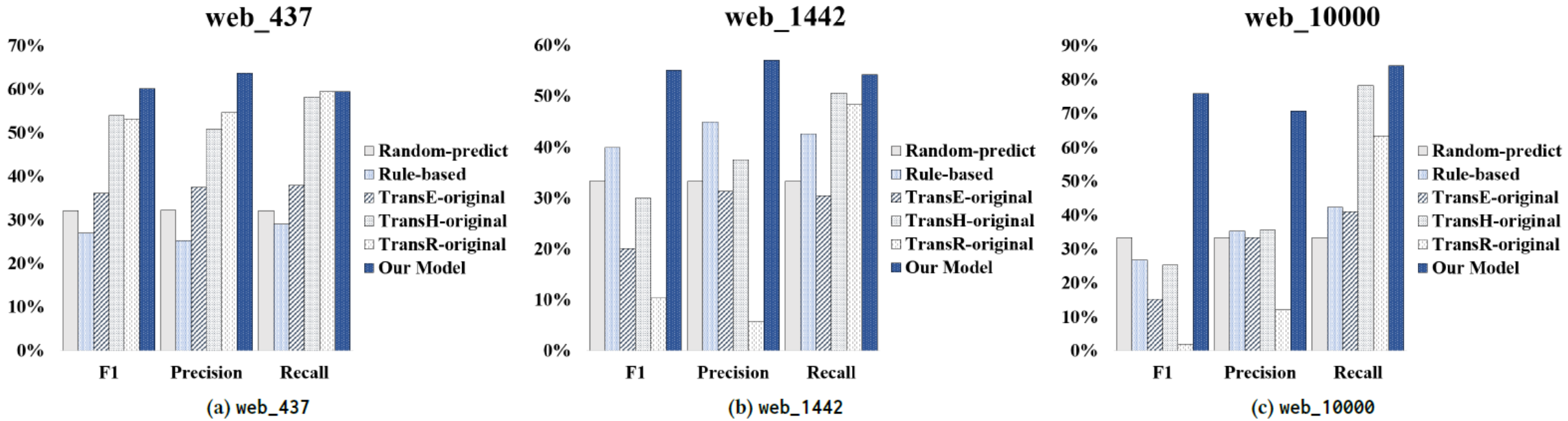
Result

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

		navigation	*suggestion*	*action*	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.



(a) web_437 (b) web_1442 (c) web_10000

- **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.

		navigation			*suggestion*			*action*	
		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