Hyperlink Classification via Structured Graph Embedding

**Main Contributions**
- We formally define a hyperlink classification problem in web search by classifying hyperlinks into three classes based on their roles: navigation, suggestion, and action.
- We approach the problem from a structured graph embedding perspective, by modifying knowledge graph embedding techniques.
- Relation perturbation in negative sampling enables us to significantly improve performance in classifying hyperlinks on web graphs.

**Real-World Web Graphs**
- The hyperlinks are created for different reasons, and play different roles.
  - Navigation links are designed to navigate the main website.
  - Suggestion links suggest users to take a look at related information.
  - Action links are made to invoke actions such as ‘edit’, or ‘send an email’.
- We create three real-world web graphs by crawling a set of web pages and the hyperlinks starting from a web page in Stack Overflow.

| $|V|$ | $|E|$ | navigation | suggestion | action |
|---|---|---|---|---|
| 437 | 404 | 437 | 268 (61.33%) | 112 (25.63%) | 57 (13.04%) |
| 10000 | 332 | 1442 | 1,284 (89.04%) | 93 (6.45%) | 65 (4.51%) |
| 2,202 | 2,010 | 10,992 (98.92%) | 85 (0.85%) | 23 (0.23%) |

All the datasets/codes are available on http://bigdata.cs.ksku.edu.

**Knowledge Graph Embedding**
- A knowledge graph is a graphical representation of human knowledge.
  - Each fact can be described as a triplet (head entity, relation, tail entity).
  - The goal of knowledge graph embedding is to represent entities and relations in a feature space while preserving the structure of the graph.
  - Given a set of golden triplets (denoted by $S$) and a set of corrupted triplets (denoted by $S'$), minimize the following loss function:

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

where $|x|_\gamma = \max(0, \gamma x)$ and $\gamma$ is the margin.

- How to compute $f(h,r,t)$ determines different embedding models.
- Given a directed web graph $G = (V, E)$ where $V = \{p_1, p_2, \cdots, p_n\}$ and $E = \{(p_i, p_j) : p_i, p_j \in V\}$, each hyperlink $r$ belongs to one of the three relation labels $R = \{n, s, a\}$.
- Given a golden triplet $(p_i, r, p_j)$, generate a corrupted triplet $(p_i, r, p_j)$.
  - Minimize the following loss function:

$$L = \sum_{(r,t,s) \in S} |f(p_i, r, p_j) - f(p_i', r, p_j')|.$$

- TransE, TransH, and TransR only corrupt entities.
- We corrupt an entity with probability $\alpha$, and corrupt the relation with probability $1 - \alpha$ ($0 < \alpha \leq 1$).

$$c(p_i, r, p_j) = \begin{cases} \text{prob. } \alpha/2 : & (p_i, r, q) \in V \setminus \{p_j\}, (p_i, r, q) \notin S, \\ \text{prob. } (1 - \alpha) : & (p_i, r, q) \in S, \end{cases}$$

For a directed edge $(p_i, r, p_j)$, we predict the relation $r'$ for $(p_i, r, p_j)$ by computing

$$r' = \text{argmin}_{r \in R} f(p_i, r, p_j)$$

where $r'$ is the predicted relation.

**Hyperlink Classification Model (Cont’d)**
- False negative: when we corrupt entities, there is a chance that it is not a corrupted one but just unobserved one in the train set.
  - If we corrupt a golden triplet $(p_i, n, p_j)$ to $(p_i, n, p_j)$, there is a risk that $(p_i, n, p_j)$ does not exist in the train set, but exist in the valid or test sets.
  - The navigation links are prevalent while there are very few suggestion and action links. This bias makes the entity corruption undesirable.
- If we corrupt a relation, it is guaranteed that the corrupted triplet is not in the test set because each pair of web pages has a unique relation.
  - If $(p_i, n, p_j)$ is observed, $(p_i, s, p_j)$ or $(p_i, a, p_j)$ should not hold.
- If we only corrupt relations and do not corrupt entities to create the negative triplets, we might have an overfitting problem and the model is not sufficiently trained for an unobserved entity.

**Experimental Results**
- The average F1 of our model with different $\alpha$ values.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\alpha = 0.3$</th>
<th>$\alpha = 0.5$</th>
<th>$\alpha = 0.7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TransE</td>
<td>34.29</td>
<td>25.39</td>
<td>33.88</td>
</tr>
<tr>
<td>TransH</td>
<td>60.25</td>
<td>58.91</td>
<td>59.83</td>
</tr>
<tr>
<td>TransR</td>
<td>57.99</td>
<td>57.32</td>
<td>54.04</td>
</tr>
</tbody>
</table>

**Conclusion & Future Work**
- By introducing an effective relation perturbation in embedding models, we can successfully classify hyperlinks on web graphs.
- We plan to extend our analysis to a case where we can incorporate various features or attributes of web pages or hyperlinks.

---