Link prediction analysis in the Wikipedia collaboration graph

Ferenc Molnár
Rensselaer Polytechnic Institute
12/06/2011
Contents

- Wikipedia collaboration graph
- Link prediction
- Analysis of predictors
- Results
Collaboration graph

- Wikipedia editors collaborate
  - editing the same page together in a timeframe
  - input: bipartite graph (editors, pages edited) in time

- Dynamic graph
  - nodes: editors
  - link strength \(\sim\) social link between editors

- Update rules
  - time resolution: weeks (same as input)
  - link strength +8 when 2 editors edit together
  - link strength -1 decay
Link prediction

- The problem
  - given the present, predict future social link formation

- Predictors, thresholds
  - principle of triadic closure
  - binary predictors
  - social link: link strength > link threshold
  - prediction: prediction value > prediction threshold

- My Project
  - analyze performance of many predictors
  - find the best one for Wikipedia collaboration graph
Analysis of predictors

- **Concept**
  - prediction ~ signal
  - reality ~ response

- **Statistics**
  - Sensitivity: $\Pr(\text{+} \mid \text{link will form})$
  - Specificity: $\Pr(\text{-} \mid \text{link will not form})$
  - Precision: $\Pr(\text{link will form} \mid \text{+})$
  - NPV: $\Pr(\text{link will not form} \mid \text{-})$
  - Accuracy: $\Pr(\text{correct prediction})$
  - F1 score:
    \[
    F_1 = 2 \frac{\text{precision} \cdot \text{sensitivity}}{\text{precision} + \text{sensitivity}}
    \]
- Adamic/Adar predictor
  - unique common neighbors are more valuable

\[
score(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log|\Gamma(z)|}
\]

- Common neighbors
  - unweighted triadic closure

\[ \text{score}(x, y) = |\Gamma(x) \cap \Gamma(y)| \]
Weighted common neighbors

depends on existing link strength

\[
score(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} S(x, z)S(z, y)
\]
- Jaccard’s coefficient
  - how specific is the common neighborhood

\[
\text{score}(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}
\]
- Preferential attachment
- rich gets richer

\[ \text{score}(x, y) = |\Gamma(x)| \cdot |\Gamma(y)| \]

**note:** smaller sample size due to computational complexity
Comparison of predictors

- Weighted common neighbors
- Adamic/Adar
- Common neighbors
- Preferential attachment
- Jaccard’s coefficient

Maximum sensitivity / specificity

Maximum accuracy

*
Comparision of predictors

- Adamic/Adar: 0.6
- Weighted common neighbors: 0.6
- Common neighbors: 0.6
- Preferential attachment: 0.55
- Jaccard's coefficient: 0.4
Conclusions

- Non-significant parameters
  - link threshold
  - prediction $\Delta T$

- Significant parameters
  - prediction threshold
  - the predictor itself

- Best predictor
  - for screening: Weighted common neighbors
  - for predicting: Weighted common neighbors, Adamic/Adar
  - for both: Adamic/Adar