Features used in “Collective Graph Identification”

1 Introduction

This document is supplementary material for Namata et al.[2] and contains detailed descriptions of the features used by the models in the experimental evaluation in the paper.

2 Cora and Citeseer

2.1 Data Description

We define the following domains:

- The set of words W=w$_1$, w$_2$, ..., w$_J$
- The set of document labels L=l$_1$, l$_2$, ..., l$_K$

We define the following predicates:

- HasWord(X, w_i): Paper X contains a word w_i
- Coref(X, Y): Papers X and Y are co-referent
- Cites(X, Y): Paper X cites paper Y
- HasLabel(X, l_i): A paper X has the label (i.e., paper topic) of l_i

In addition, in some places, we make a distinction between CitesObs(X,Y), which is a citation link that is observed in the input, and CitesInf(X,Y), which is a citation link that is inferred from the input. Cites(X,Y) is the union of CitesObs(X,Y) and CitesInf(X,Y).

In the following sections, we define the feature functions (highlighted in bold) used, along with auxilary function definitions. For entity resolution and link prediction, the functions take two arguments corresponding to the nodes of the potential co-referent pair or potential link. Node labeling takes one argument corresponding to the node.

2.2 Entity Resolution

We define the following feature functions:

- WordOccurences(X) = the set of words that appear in document X
- **WordJaccard(X, Y)** = $\frac{|\text{WordOccurrence(X)}\cap\text{WordOccurrence(Y)}|}{|\text{WordOccurrence(X)}\cup\text{WordOccurrence(Y)}|}$
- ObsNeighbors(X) = {Y | CitesObs(X, Y)}, the set of nodes that have observed citation edges with X
- **ObsNeighborJaccard(X, Y)** = $\frac{|\text{ObsNeighbors(X)}\cap\text{ObsNeighbors(Y)}|}{|\text{ObsNeighbors(X)}\cup\text{ObsNeighbors(Y)}|}$
- CitesNeighbors(X) = {Y | Cites(X,Y)}, the set of nodes that have observed or inferred citation edges with X
- **CitesNeighborJaccard(X, Y)** = $\frac{|\text{CitesNeighbors(X)}\cap\text{CitesNeighbors(Y)}|}{|\text{CitesNeighbors(X)}\cup\text{CitesNeighbors(Y)}|}$
- Coref*(X, Y) = {(X, Y) | Coref(X, Y) $\vee$ (Coref(X, Z) $\wedge$ Coref*(Z, Y))}, the transitive closure of Coref
- CorefObsNeighbors(X) = {Z | CitesObs(X, Y) $\wedge$ Coref*(Y, Z)}, the set of nodes that are co-referent to nodes which have observed citation edges with X
- **CorefObsNeighborJaccard(X, Y)** = $\frac{|\text{CorefObsNeighbors(X)}\cap\text{CorefObsNeighbors(Y)}|}{|\text{CorefObsNeighbors(X)}\cup\text{CorefObsNeighbors(Y)}|}$
- CorefCitesNeighbors(X) = {Z | (Cites(X, Y)) $\wedge$ Coref*(Y, Z)}, the set of nodes that are co-referent to nodes which have inferred or observed citation edges with X
- **CorefCitesNeighborJaccard(X, Y)** = $\frac{|\text{CorefCitesNeighbors(X)}\cap\text{CorefCitesNeighbors(Y)}|}{|\text{CorefCitesNeighbors(X)}\cup\text{CorefCitesNeighbors(Y)}|}$
- **HasLabel-**l_i, l_j**(X, Y)** = HasLabel(X, l_i) $\wedge$ HasLabel(Y, l_j), for each possible label pair l_i, l_j
- **IsCoref(X,Y)** = 1 iff Coref*(X, Y) is non-empty, 0 otherwise

2.3 Link Prediction

In addition to using the features defined above, **WordJaccard(X, Y)**, **HasLabel-**l_i, l_j**(X,Y)**, we define the following features for link prediction:

- **SameWord-**w_i**(X,Y)** = HasWord(X, w_i) $\wedge$ HasWord(Y, w_i) for each word w_i
- **LinkedCorefs(X, Y)** = Coref*(X, A) $\wedge$ Coref*(X, B) $\wedge$ Cites(A, B)

- **HasCommonObsNeighbor(X, Y)** = ∃Z(CitesObs(X, Z) ∧ CitesObs(Y, Z))
- **HasCommonCitesNeighbor(X, Y)** = ∃Z(Cites(X, Z) ∧ Cites(Y, Z))

2.4 Collective Classification

In addition to using the feature defined above, **HasWord(X,** w_i), we define the following features for collective classification:

- ObsNeighbors-l_i(X) = {Y | CitesObs(X, Y) ∧ HasLabel(Y, l_i)}, the set of nodes with label l_i that have observed citation edges with X, for each l_i
- **ObsNeighborProportion-**l_i**(X)** = $\frac{|\text{ObsNeighbors-}l_i(\text{X})|}{|\text{ObsNeighbors(X)}|}$, for each l_i
- CitesNeighbors-l_i(X) = {Y | Cites(X, Y) ∧ HasLabel(Y, l_i)}, the set of nodes with label l_i that have inferred or observed citation edges with X, for each l_i
- **CitesNeighborProportion-**l_i**(X)** = $\frac{|\text{CitesNeighbors-}l_i(\text{X})|}{|\text{CitesNeighbors(X)}|}$, for each l_i
- Coref*(X) = {Y | Coref*(X, Y)}, the set of nodes co-referent, by transitive closure, to X
- Coref*-l_i(X) = {Y | Coref*(X, Y) ∧ HasLabel(Y, l_i)} the set of nodes with label l_i co-referent, by transitive closure, to X, for each l_i
- **CorefCitesNeighborProportion-**l_i**(X)** = $\frac{|(\text{Coref}^*\text{-}l_i(\text{X}))|}{|\text{Coref}^*(\text{X})|}$, for each l_i

3 Enron Communication Network

3.1 Data Description

We define the following domains:

- The set of words W=w_1, w_2, ..., w_J
- The set of person labels L=l_1, l_2, ..., l_K

We define the following predicates:

- HasAddress(X, Y): Email account X has the email address Y
- HasWord(X, w_i): Email account X has used word Y at least once in its communications
- Communicates(X, Y): Email accounts X and Y have shared at least one communication

- Manages(X, Y): User of email accounts X and Y share a managerial relationship
- Communication(C, X, Y): Communication C was sent from X to Y
- CommunicationHasWord(X, w_i): Communication X contains word w_i
- Coref(X, Y): Email account X and Y are owned by the same person
- HasLabel(X, l_i): A person X has the label (i.e., title) of l_i

3.2 Entity Resolution

We define the following feature functions:

- WordOccurences(X) = the set of words that appear in email account X
- **WordJaccard(X, Y)** = $\frac{|\text{WordOccurrence(X)} \cap \text{WordOccurrence(Y)}|}{|\text{WordOccurrence(X)} \cup \text{WordOccurrence(Y)}|}$
- **AddressSimilarity(X, Y)** = Q-Gram[1] string similarity of the email addresses of X and Y
- CommNeighbors(X) = {Y | Communicates(X, Y)}, the set of nodes that have communication edges with X
- **CommNeighborJaccard(X, Y)** = $\frac{|\text{CommNeighbors(X)} \cap \text{CommNeighbors(Y)}|}{|\text{CommNeighbors(X)} \cup \text{CommNeighbors(Y)}|}$
- Coref*(X,Y) = {(X, Y) | Coref(X, Y) $\vee$ (Coref(X, Z) $\wedge$ Coref*(Z, Y))}, the transitive closure of Coref
- CorefCommNeighbors(X) = {Z | Communicates(X, Y) $\wedge$ Coref*(Y,Z)}, the set of nodes that are co-referent to nodes which have an observed communication edge with X
- **CorefCommNeighborJaccard(X, Y)** = $\frac{|\text{CorefCommNeighbors(X)} \cap \text{CorefCommNeighbors(Y)}|}{|\text{CorefCommNeighbors(X)} \cup \text{CorefCommNeighbors(Y)}|}$
- CorefMngNeighbors(X) = {Z | (Manages(X, Y)) $\wedge$ Coref*(Y, Z)}, the set of nodes that are co-referent to nodes which have inferred or observed managerial edges with X
- **CorefMngNeighborJaccard(X, Y)** = $\frac{|\text{CorefMngNeighbors(X)} \cap \text{CorefMngNeighbors(Y)}|}{|\text{CorefMngNeighbors(X)} \cup \text{CorefMngNeighbors(Y)}|}$
- **HasLabel-**l_i, l_j**(X,Y)** = HasLabel(X, l_i)$\wedge$HasLabel(Y, l_j), for each possible label pair l_i, l_j
- **IsCoref(X,Y)** = 1 iff Coref*(X, Y) is non-empty, 0 otherwise

3.3 Link Prediction

In addition to using the features defined above, **WordJaccard(X, Y)** and **HasLabel-**l_i, l_j**(X, Y)**, we define the following features for link prediction:

- **SharedWord-**w_i**(X, Y)** = $\exists Z$(Communication(C, X, Y)$\wedge$CommunicationHasWord(C, w_i))
- Communications(X, Y) = {C | Communication(C, X, Y)}, the set of communications between X and Y
- **CommunicationCount(X, Y)** = |Communications(X)|
- **LinkedCorefs(X, Y)** = Coref*(X, A) $\wedge$ Coref*(X, B) $\wedge$ Communicates(A, B)
- **IsInCommMng(X, Y)** = $\exists C$(Communication(C, Z, X) $\wedge$ Manages(Z, Y))
- **IsOutCommMng(X, Y)** = $\exists C$(Communication(C, X, Z) $\wedge$ Manages(Z, Y))
- **IsCommMng(X, Y)** = $\exists C$(Communicates(X, Z) $\wedge$ Manages(Z, Y))

3.4 Collective Classification

In addition to using the feature defined above, **HasWord(X,** w_i**)**, we define the following features for collective classification:

- SentCommunications(X) = {C | Communication(C, X, Y)}, the set of communications sent by X
- **NumSent(X)** = |SentCommunications(X)|
- ReceivedCommunications(X) = {C | Communication(C, Y, X)}, the set of communications received by X
- **NumReceived(X)** = |ReceivedCommunications(X)|
- **NumSentOrReceived(X)** = |SentCommunications(X)|+|ReceivedCommunications(X)|
- InCommNeighbors(X) = {Y | Communication(C, Y, X)}, the set of nodes that have a communication sent to X
- InCommNeighbors-l_i(X) = {Y | Communication(C, Y, X)}, the set of nodes with label l_i that have a communication sent to X, for each l_i
- **InCommNeighborProportion-**l_i**(X)** = $\frac{|\text{InCommNeighbors-}l_i(\text{X})|}{|\text{InCommNeighbors(X)}|}$, for each l_i
- OutCommNeighbors(X) = {Y | Communication(C, X, Y)}, the set of nodes that have a communication from X
- OutCommNeighbors-l_i(X) = {Y | Communication(C, X, Y)}, the set of nodes with label l_i that have a communication from X, for each l_i

- **OutCommNeighborProportion-l_i(X)** = $\frac{|\text{OutCommNeighbors-}l_i(\text{X})|}{|\text{OutCommNeighbors(X)}|}$, for each l_i
- CommNeighbors-l_i(X) = {Y | Communicates(X, Y) ∧ HasLabel(Y, l_i)}, the set of nodes with label l_i that have communication edges with X, for each l_i
- **CommNeighborProportion-l_i(X)** = $\frac{|\text{CommNeighbors-}l_i(\text{X})|}{|\text{CommNeighbors(X)}|}$, for each l_i
- InComm(X) = {C | Communication(C, Y, X)}, the set of communications to X
- InComm-l_i(X) = {Y | Communication(C, Y, X)}, the set of communications to X from nodes with label l_i, for each l_i
- **InCommProportion-l_i(X)** = $\frac{|\text{InComm-}l_i(\text{X})|}{|\text{InComm(X)}|}$, for each l_i
- OutComm(X) = {C | Communication(C, X, Y)}, the set of communications from X
- OutComm-l_i(X) = {Y | Communication(C, X, Y)}, the set of communications from X to nodes with label l_i, for each l_i
- **OutCommProportion-l_i(X)** = $\frac{|\text{OutComm-}l_i(\text{X})|}{|\text{OutComm(X)}|}$, for each l_i
- **InAndOutCommProportion-l_i(X)** = $\frac{|\text{InComm-}l_i(\text{X})+\text{OutComm-}l_i(\text{X})|}{|\text{InComm(X)+OutComm(X)}|}$, for each l_i
- MngNeighbors(X) = {Y | Manages(X, Y)}, the set of nodes that have inferred or observed managerial edges with X
- MngNeighbors-l_i(X) = {Y | Manages(X, Y) ∧ HasLabel(Y, l_i)}, the set of nodes with label l_i that have inferred or observed managerial edges with X, for each l_i
- **MngNeighborProportion-l_i(X)** = $\frac{|\text{MngNeighbors-}l_i(\text{X})|}{|\text{MngNeighbors(X)}|}$, for each l_i
- Coref*(X) = {Y | Coref*(X, Y)}, the set of nodes co-referent, by transitive closure, to X
- Coref*-l_i(X) = {Y | Coref*(X, Y) ∧ HasLabel(Y, l_i)} the set of nodes with label l_i co-referent, by transitive closure, to X, for each l_i
- **CorefCitesNeighborProportion-l_i(X)** = $\frac{|(\text{Coref}^*\text{-}l_i(\text{X}))|}{|\text{Coref}^*(\text{X})|}$, for each l_i

References

[1] L. Gravano, P. G. Ipeirotis, H. V. Jagadish, N. Koudas, S. Muthukrishnan, L. Pietarinen, and D. Srivastava. Using q-grams in a dbms for approximate string processing. *IEEE Data Engineering Bulletin*, 24:28–34, 2001.

[2] G. M. Namata, S. Kok, and L. Getoor. Collective graph identification. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011.