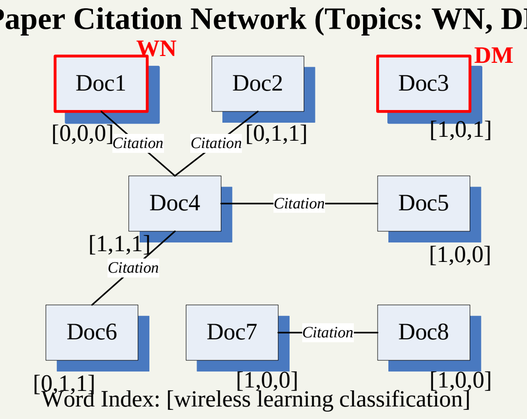
Combining Link and Content for Collective Active Learning

Lixin Shi, Yuhang Zhao, and Jie Tang
Dept. of Computer Science and Technology, Tsinghua University
{shilixinhere, zhaoyh630}@gmail.com, tangjie@keg.cs.tsinghua.edu.cn

Arnetminer
Http://www.arnetminer.org

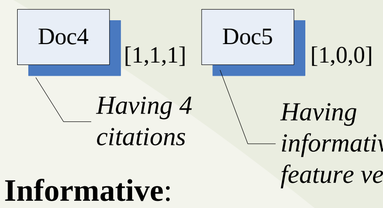
Paper Citation Network (Topics: WN, DM)



Word Index: [wireless learning classification]

Select *k* most informative samples from unlabeled set

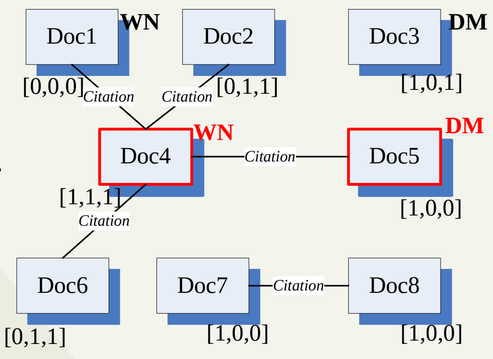
***k* = 2:**



Informative: combined both **content** and **link**

Query the User

Paper Citation Network (Topics: WN, DM)



Word Index: [wireless learning classification]

Collective Classifier

Random Walk Framework

Table of symbols

Symbol	Description
$\mathcal{U}$	The set of unlabeled data instances
$\mathcal{L}$	The set of labeled data instances
$\mathcal{G}$	The network representing the relationships between instances
$\boldsymbol{x}_i$	The feature vector of instance i
y_i	The label of instance $i \in \mathcal{L}$
W	The similarity matrix
P	The transition matrix in the random walk framework
$\boldsymbol{f}_u$	Expectation vector of instances in unlabeled data set
$\boldsymbol{f}_l$	Expectation vector of instances in labeled data set
f_i	The expectation of instance i's label
S	The set of instances to be selected
k	The number of instances to be selected

Calculating transition Probability

link: $l_{ij} = \frac{1}{d_i}$, Where d_j is number of links going out of $\boldsymbol{x}_i$

Similarity: $w_{ij} = \exp\left(-\frac{1}{\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right)$

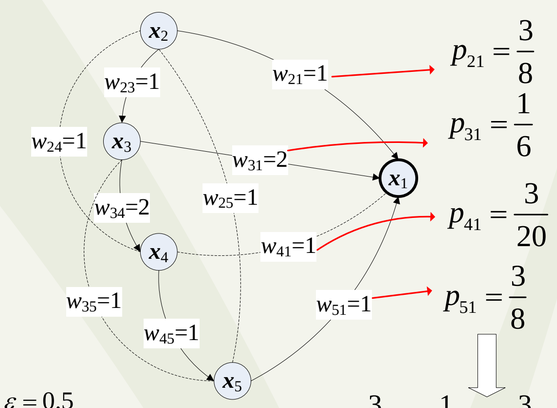
Transition Probability: $p_{ij} = \varepsilon l_{ij} + (1-\varepsilon)\frac{w_{ij}}{\sum_k w_{ik}}$ (Link / Content)

$\vec{f} = P\vec{f}$, where $\vec{f} = [f_1, f_2, \ldots, f_n]^T, P = (p_{ij})_{n\times n}$

f_i: expectation of label of instance i

p_{ij}: Transition Probability from i to j

An Example



Our Approach

$$Q(S) = \alpha C(S) + (1-\alpha)H(S), 0 \le \alpha \le 1, \text{where } H(S) = \sum_{i\in S} H(i) = \sum_{i\in S} f_i \log\frac{1}{f_i} + (1-f_i)\log\frac{1}{1-f_i}$$

$$C(S) = \sum_{i\in U}(H(i))^{\beta}\left(\max_{j\in L\cup S} w_{ij}\right)^{1-\beta} = \sum_{i\in U}\left(f_i \log\frac{1}{f_i} + (1-f_i)\log\frac{1}{1-f_i}\right)^{\beta}\left(\max_{j\in L\cup S} w_{ij}\right)^{1-\beta}$$

We designed $Q(S)$ as the objective function over S, a subset of the unlabeled set. It measures the informativeness of S, so collective active learning could be viewed as selecting:

$$S = \arg\max_{S\in U, |S|\le k}\{Q(S)\}$$

Criteria of Objective Function

We use three criteria to design $Q(S)$:

- Maximum Uncertainty: $H(S)$ as summation of entropies
- Maximum Impact: Graph-cut-based design of $C(S)$, which could be seen as a summation of maximum impacts over samples in S
- Minimum Redundancy: We theoretically proved that definition of $C(S)$ minimizes redundancy

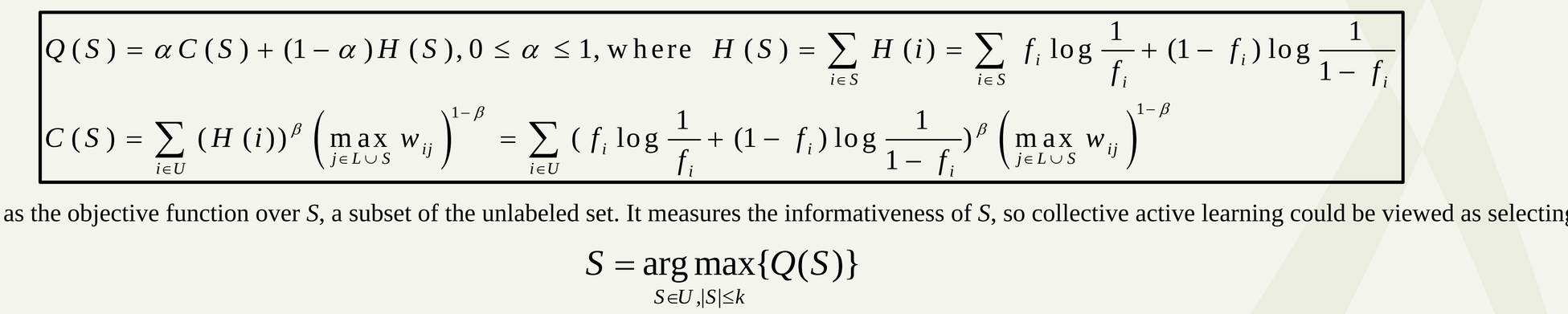
Algorithm Design

- Submodularity: Our definition of Q(S) satisfy the monotonic submodularity property, which guarantees a greedy algorithm
- Error Bound: This greedy algorithm has an approximation rate of (1-1/*e*)
- Speedups: We parallelized the algorithm for scaling up to real large data sets

Experiments

Synthetic Data Set:

- Two classes denoted by different colors
- +: initially labeled samples
- O: selected samples
- Demonstrate the necessity of all three criteria

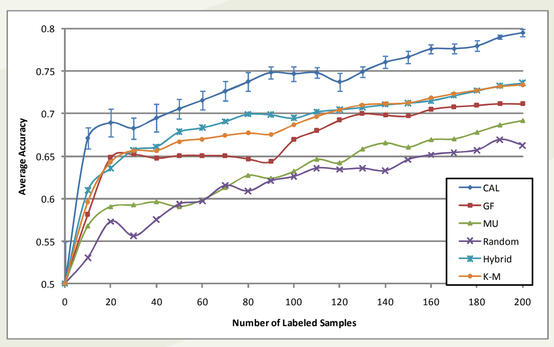


Proposed(80.25%) | **Q=H(S) (76.93%)** | **Q=C(S) (75.74%)** | **Without Maximum Redundancy (80.25%)**

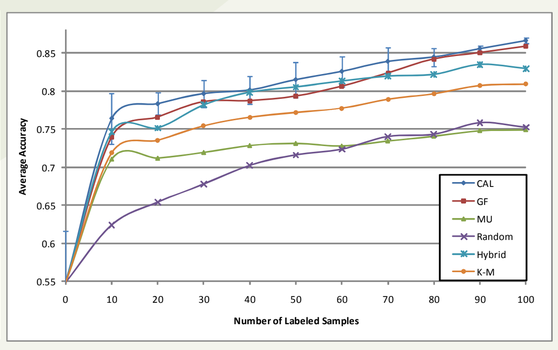
Networked Data Sets:

- 2 text classification data sets, with citation links, 1 with web links
- Outperform up to 6% compared with State-of-the-art Baselines
- Stable performance over different sets

Citeseer Data Set



WebKB Data Set



Cora Data Set