Extracting knowledge from graph structure
(10-805 project proposal)

Jakub Pachocki

March 11, 2015

1 Idea

Imagine we are given a graph over concepts, with edges between related concepts. In particular, take Wikipedia as a simple unlabeled graph, with undirected edges between the articles corresponding to links. The basic question I would like to investigate is: how well can we classify the articles (their importance, broad category, length, etc. . .) based only on the graph structure.

2 Techniques

There are two major approaches to the problem:

- *Global* approach. Attempt to cluster (or compute eigenvectors, etc. . .) the entire graph. Use the resulting information as features to perform classification.
- *Local* approach. Assume when classifying we only have access to an anonymized version of the neighborhood of a node. What sort of local features can we compute to help with the classification? (Degree, density of neighborhood, number of triangles in neighborhood, etc. . .)

3 Desired results and prior work

I would like to investigate the graph structure of Wikipedia and determine how strongly it correlates with article categories. In particular I want to classify Wikipedia articles into a category hierarchy based on the spectral embedding of the hyperlink graph.

Another question of interest is deducing nontrivial things about an object by only looking at its neighborhood in the graph. There are very interesting results of this kind in social networks, for example in the analysis of Facebook neighborhood graphs (Backstrom & Kleinberg, 2013: http://arxiv.org/abs/1310.6753).

4 Dataset

I plan to focus on using the link dataset collected by DBpedia (http://wiki.dbpedia.org/).

My classification labels are given by the category hierarchy at http://mappings.dbpedia.org/server/ontology/classes/. There have been attempts to classify Wikipedia into this hierarchy based on article text (http://lshtc.iit.demokritos.gr/); they should provide a good baseline for comparison.

An interesting question that I also plan to investigate is whether mixing in graph features can improve the performance of text-based classifiers.

5 Collaborators

Currently I have no collaborators. I would be happy to work with 605 students who are interested in the idea. I anticipate there is plenty of engaging work to be done in implementing the graph algorithms and interpreting their results.