Abstract

The objective of this project is to build an automated system that suggests missing categories of Wikipedia articles to editors through a browser-based user-interface. The system learns from Wikipedia’s existing corpus of human-categorized articles and provides a ranked list of new category suggestions. All suggested categorizations are subject to final approval or rejection by a human editor.

Motivation

Wikipedia’s complex structure makes relevant categories difficult for users to find:

- There is too much work and human error involved in categorizing a page.
- Wikipedia’s current human-edited category system lends itself well to algorithmic analysis and improvement.

System Design

The graph algorithm generates suggestions by looking at pages with shared categories and sampling the categories of those pages.

Graph Algorithm

Add categories to Wikipedia

Score Suggestions

Filtering Data and Graph Algorithm

WikiCat

Interface

Editors are presented with a simple interface for easily adding our suggested categories to articles.

Evaluation:

- Is it making good suggestions?
- Is it making bad suggestions?
- WikiCat improves as editors use it.

Goals

- Utilize both graph-based techniques and machine learning to develop an algorithm which suggests missing categories for Wikipedia articles.
- Integrate the results of the algorithm into existing Wikipedia editing tool HotCat.
- Deploy tool to actual editors and collect feedback.

System Design

Annotator Interface

Top N suggested suggestions

Human Labeling

Wikipedia Editor

WikiCat

Graph Algorithm

Articles

Links

Categories

Top N: Article

Category

Score

Feature Determination

Tuples: Article

Category

Features

Add categories to Wikipedia

Score Suggestions

Filtering Data and Graph Algorithm

WikiCat