

---

# EECS 405 Project Proposal

for

Improving Efficiency of String Similarity Searches  
through Clustered Pruning

Prepared by

James Fitzpatrick  
Kyle Patterson

Version 1.0  
March 18, 2014

---

### Revision History

Name	Date	Reasons for Change	Version
James Fitzpatrick	3/18/14	Initial Proposal	

## **Contents**

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Problem Definition</b>	<b>3</b>
<b>3</b>	<b>Goals</b>	<b>3</b>
<b>4</b>	<b>Division of Labor</b>	<b>4</b>
<b>5</b>	<b>Papers Identified</b>	<b>4</b>

## 1. Introduction

## 2. Problem Definition

We propose a method of pruning the search space that utilizes a common data mining technique, clustering. At the application's initialization, we will mine the database and identify similar groups of strings; these clusters will be indexed by longest common substring. When a search is performed at run time, we will search for the appropriate cluster of strings using the constructed cluster index and restrict our search space to the appropriate cluster of data. Although this method includes a significant amount of preprocessing, index construction is only needed when the data is modified and is not present in every search. This effectively reduces the number of strings that must be compared via the  $B^{\text{ed}}$ -Tree and Top- $k$  techniques from [1] and [2], which should further improve search time among large sets of data.

As presented in [9], using longest common subsequence (LCS) is the second most effective technique in grouping strings (with the Levenshtein distance being the most effective). However, LCS is more conducive to key generation because it provides a testable string that can be used as the key. The best-matching LCS of a query and a key is guaranteed to identify the cluster that contains the query, if the query is present in the database.

## 3. Goals

1. Research and gain a thorough understanding of the research papers
2. Implement the  $B^{\text{ed}}$  Tree and Top-K Algorithms
3. Build clustered datasets based on the Longest Common Subsequence on:
  - Movie Titles
  - Publication Titles
4. Test the effects on the performance of the proposed clustering techniques

5. Formally present findings through a final presentation

## 4. Division of Labor

James Fitzpatrick will be implementing the Top-K Algorithm as documented in the papers below. James will also create an agnostic test harness to assess the differences between the different search algorithms and the different clustering techniques on the datasets.

Kyle Patterson will be implementing the B<sup>ed</sup> Tree Algorithm and will be developing the clustering algorithms for the datasets.

## 5. Papers Identified

1. Z. Zhang, et. al. "B<sup>ed</sup>-Tree: An All-Purpose Index Structure for String Similarity Search Based on Edit Distance", SIGMOD 2010.
2. Don Deng, Guoliang Li, Jianhua Feng, Wen-Syan Li, Top-*k* String Similarity Search with Edit-Distance Constraints, ICDE 2013.
3. A. Andoni and K. Onak. Approximating edit distance in near-linear time. STOC, pages 199-204, 2009.
4. M. Hadjieleftheriou, A. Chandel, N. Koudas, and D. Srivastava. Fast indexes and algorithms for set similarity selection queries. ICDE 2008, pages 267-276.
5. M. Hadjieleftheriou, N. Koudas, and D. Srivastava. Incremental maintenance of length normalized indexes for approximate string matching. *SIGMOD Conference*, pages 429-440, 2009.
6. W. J. Masek and M. Paterson. A faster algorithm computing string edit distances. *Journal of Computer and System Sciences*, 20(1):18-31, 1980.
7. T. Kahveci and A. K. Singh. Efficient index structures for string databases. VLDB, pages 351-360, 2001.

8. Z. Yang, J. Yu and M. Kitsuregawa. Fast algorithms for top- $k$  approximate string matching. AAAI, 2010.
9. R. Zafarani and H. Liu. Connecting users across social media sites: a behavioral-modeling approach. KDD, pages 41-49, 2013.
10. M. Rafsanjani, Z. Varzaneh, N. Chukanlo. A survey of hierarchical clustering algorithms. TJMCS 5(3):229-240, 2012.