Srikanth Sundaram

Avinash Ranganath

Philipp Petrenz

Web Document Clustering: A Feasibility Demonstration

O. Zamir & O. Etzioni

Overview

- Motivation — **Why Clustering of Search Results?**
- Suffix Tree Clustering — **Explaining the Algorithm**
- Evaluation — **How well can we do?**
- Live Demo — **The fun part (we promise)**

Motivation

Relevance

Browsable Summaries

Overlap

Snippet-tolerance

Speed

Incrementality

Motivation

Relevance

Browsable Summaries

Overlap

Snippet-tolerance

Speed

Incrementality

Google Query: “Salsa”

We need to produce clusters which group documents relevant to user’s query separately from irrelevant ones.

Motivation

Relevance

Browsable Summaries

Overlap

Snippet-tolerance

Speed

Incrementality

UKSalsa.com - Your guide to **Salsa** in the UK
12 Aug 2005 ... Uksalsa.com is a guide to the UK **Salsa** scene. It contains listings for **Salsa** clubs, a calendar page for special **Salsa** events, articles about ...
www.uk**salsa**.com/ - 68k - Cached - Similar pages -

SALSA - Safe and Local Supplier Approval
SALSA is a new supplier approval scheme designed to help local and regional food and drink producers supply their products to national and regional buyers.
www.**salsa**food.co.uk/ - 11k - Cached - Similar pages -

Avocado Corn **Salsa** Recipe
Avocado Corn **Salsa** Recipe, from the archives of Recipe Ideas.
www.recipe-ideas.co.uk/recipes-5/Avocado%20Corn%20**Salsa**.htm - 10k - Cached - Similar pages -

Welcome to **Salsa** Londons Number 1 Latin Live Music Venue
Welcome to **Salsa** Londons Number 1 Latin Live Music Venue, Food Served At Bar **Salsa**! Organize Parties At Bar **Salsa**! Our Clubs! Live Music At Bar **Salsa**!
www.bar**salsa**.info/ - 25k - Cached - Similar pages -

Salsa Jive UK: classes, clubs, events, news, chat and more
The Live UK What's On Guide for **Salsa** and Modern Jive (Leroc, Ceroc) - dance classes, events, workshops, news, chat and more...
www.**salsa**jive.co.uk/ - 6k - Cached - Similar pages -

Salsa & Merengue Society Homepage
Excellent homesite featuring **salsa** and merengue tutorials, **salsa** and merengue music database, **salsa** teachers course, history of **salsa**, history of merengue ...
www.**salsa**-merengue.co.uk/ - 7k - Cached - Similar pages -

We need to produce clusters which group documents relevant to user's query separately from irrelevant ones.

Motivation

Relevance

Browsable Summaries

Overlap

Snippet-tolerance

Speed

Incrementality

These phrases should provide accurate description of the clusters.

Motivation

A document might have multiple topics, it is important to allocate this document to different clusters.

Relevance

Browsable Summaries

Overlap

Snippet-tolerance

Speed

Incrementality

Motivation

For impatient users, each second counts.

Relevance

Browsable Summaries

Overlap

Speed

Snippet-tolerance

Incrementality

Motivation

Relevance

Browsable Summaries

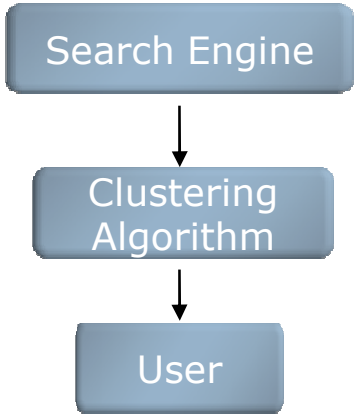
Overlap

Speed

Snippet-tolerance

Incrementality

To produce high quality clusters even when it only has access to snippets returned by search engines.



Motivation

Relevance

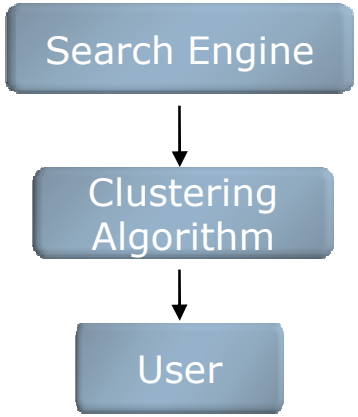
Browsable Summaries

Overlap

Speed

Snippet-tolerance

Incrementality



The method should process the documents as soon as we receive it over the web.

Motivation

Challenging existing algorithms

- Time Complexity

- Treating documents as sequence of words

- Search Engines like Google provide improvements or recommendations on query and not clustering.

Searches related to: **university of edinburgh**

university of **glasgow**	university of **aberdeen**	**napier** university	university of **dundee**
university of **strathclyde**	university of **st andrews**	university of **exeter**	**heriot watt** university

Motivation

Challenging existing algorithms

- Time Complexity

- Treating documents as sequence of words

- Search Engines like Google provide improvements or recommendations on query and not clustering.

Instead →

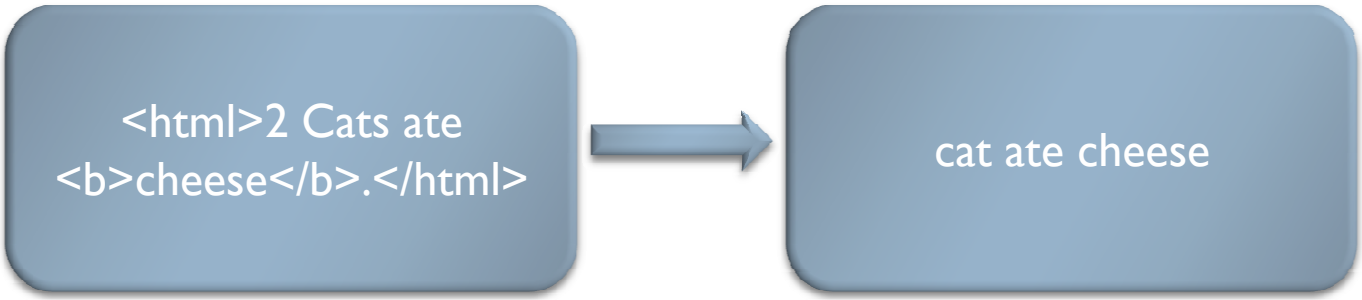
Suffix Tree Clustering

Easy as 1-2-3!

1. Clean your Documents
2. Identify your Base Clusters
3. Cluster your Clusters some more

Document Cleaning

- Stemming
- Stripping of HTML, punctuation and numbers



Identifying Base Clusters

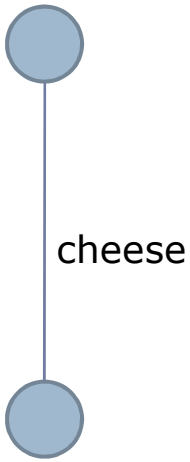
Growing our very own Suffix Tree!

1. cat ate cheese
2. mouse ate cheese too
3. cat ate mouse too

Identifying Base Clusters

Growing our very own Suffix Tree!

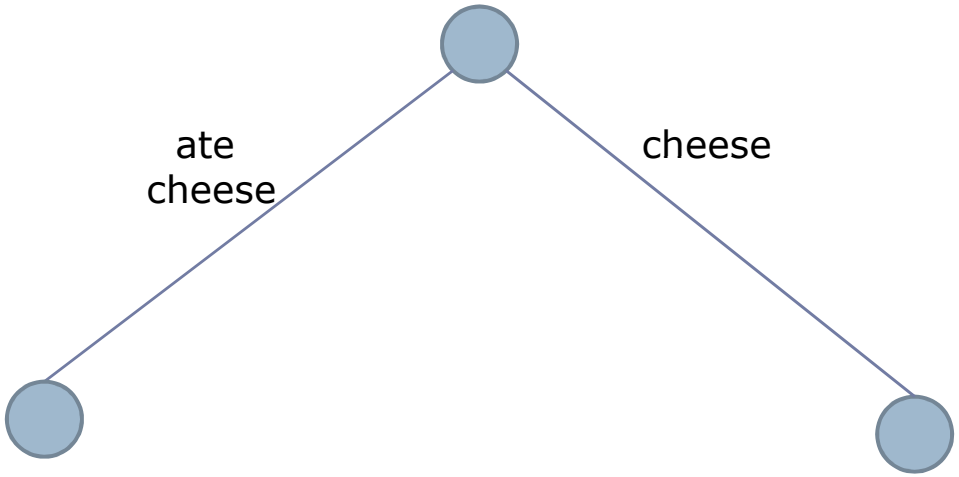
1. cat ate **cheese**
2. mouse ate cheese too
3. cat ate mouse too



Identifying Base Clusters

Growing our very own Suffix Tree!

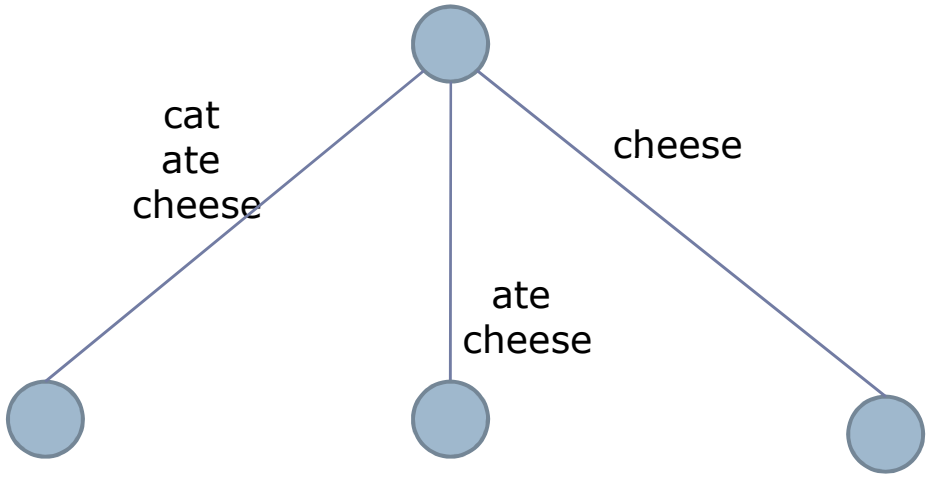
1. cat **ate cheese**
2. mouse ate cheese too
3. cat ate mouse too



Identifying Base Clusters

Growing our very own Suffix Tree!

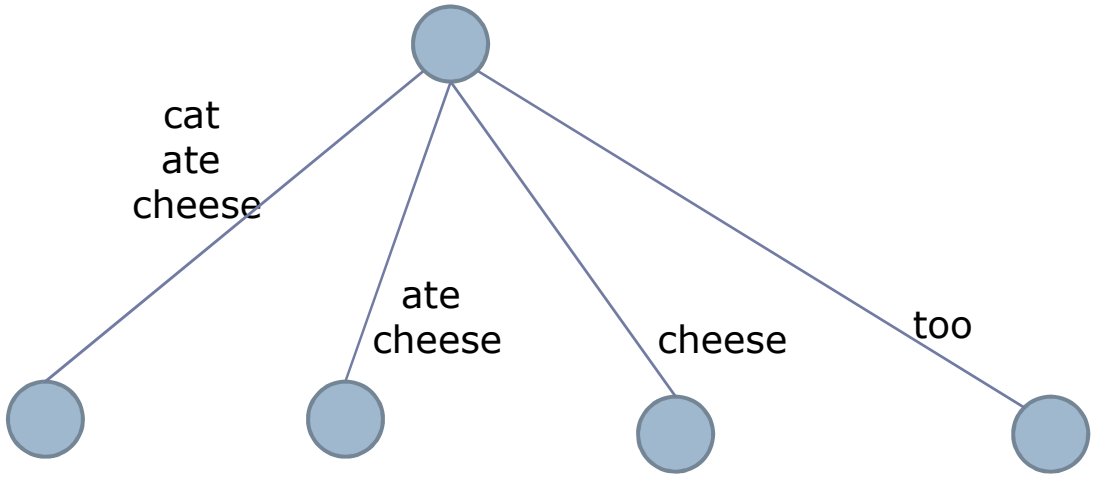
1. cat ate cheese
2. mouse ate cheese too
3. cat ate mouse too



Identifying Base Clusters

Growing our very own Suffix Tree!

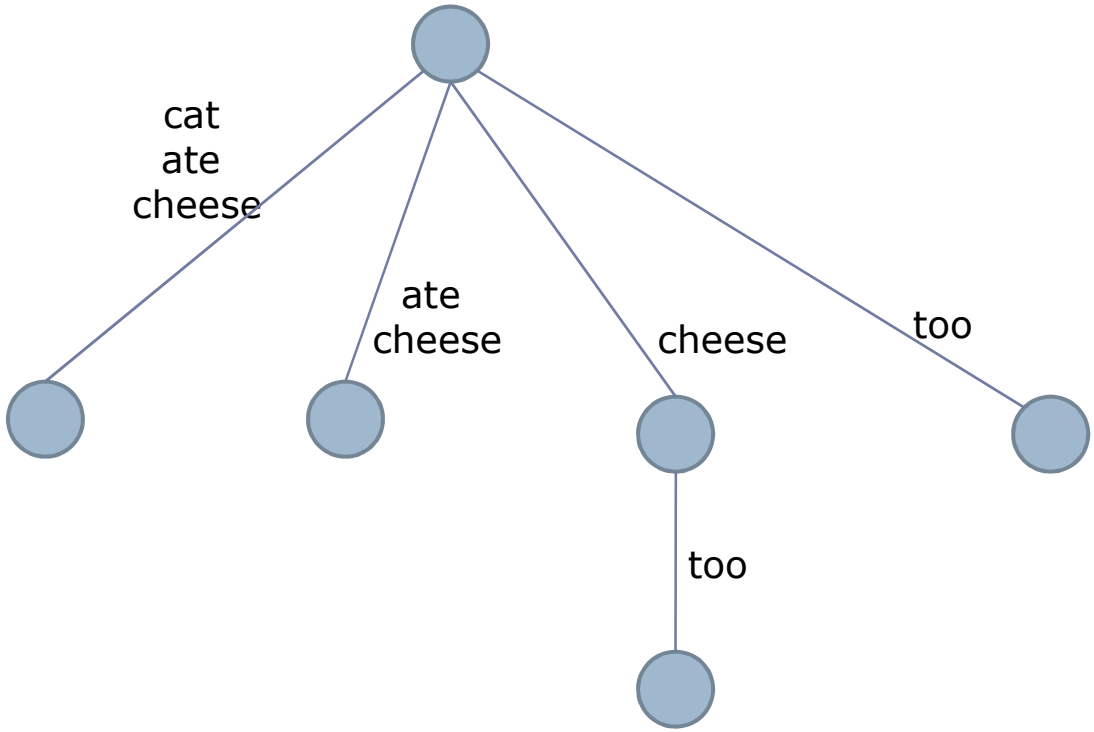
1. cat ate cheese
2. mouse ate cheese too
3. cat ate mouse too



Identifying Base Clusters

Growing our very own Suffix Tree!

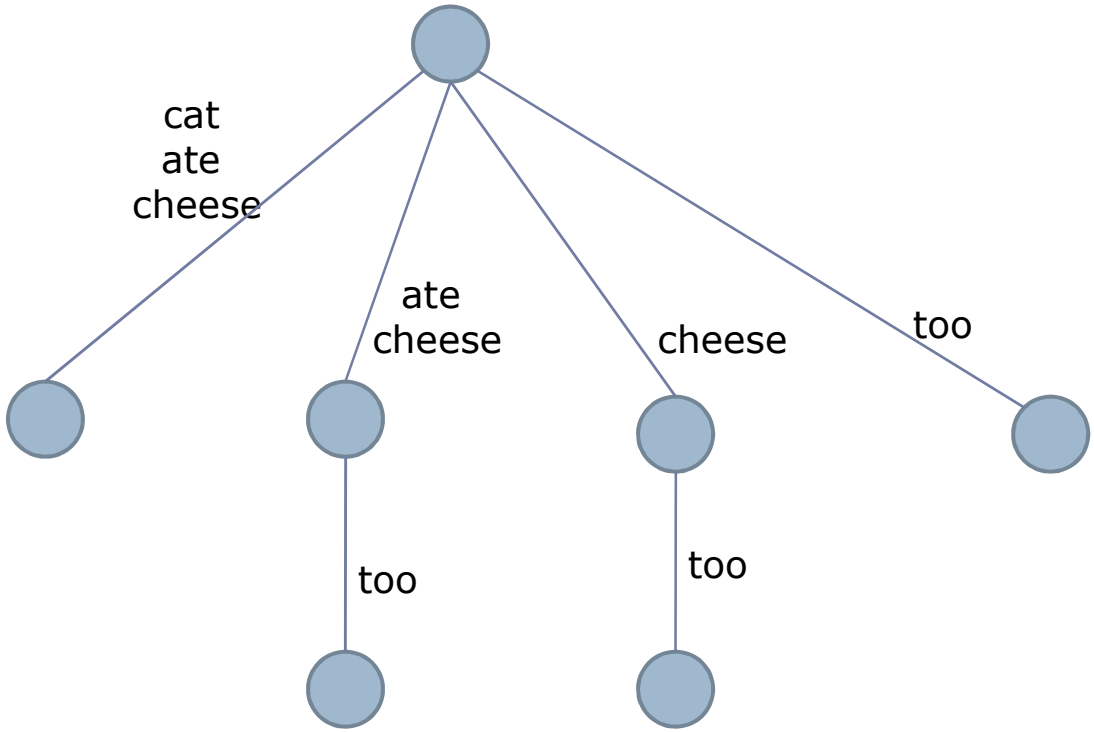
1. cat ate cheese
2. mouse ate cheese too
3. cat ate mouse too



Identifying Base Clusters

Growing our very own Suffix Tree!

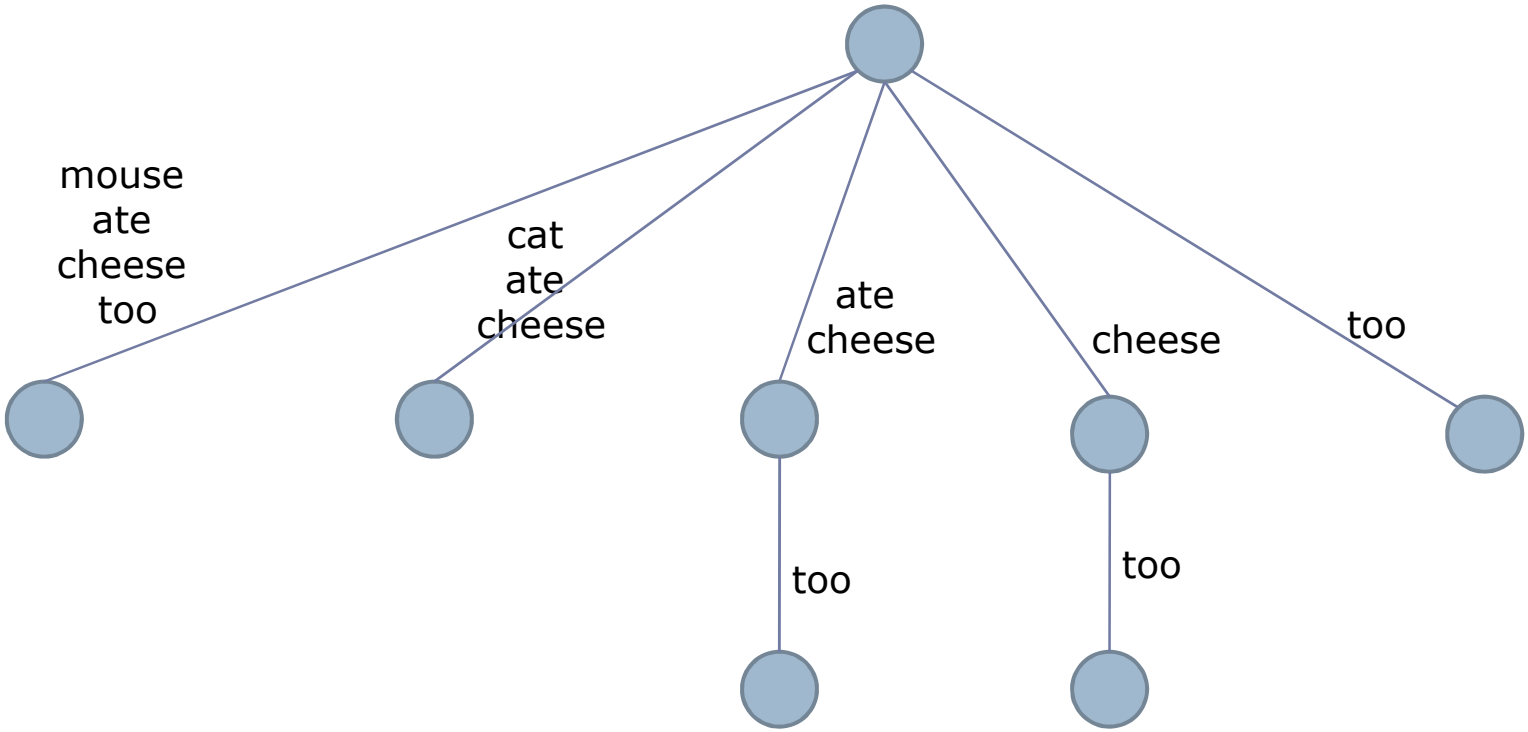
1. cat ate cheese
2. mouse ate cheese too
3. cat ate mouse too



Identifying Base Clusters

Growing our very own Suffix Tree!

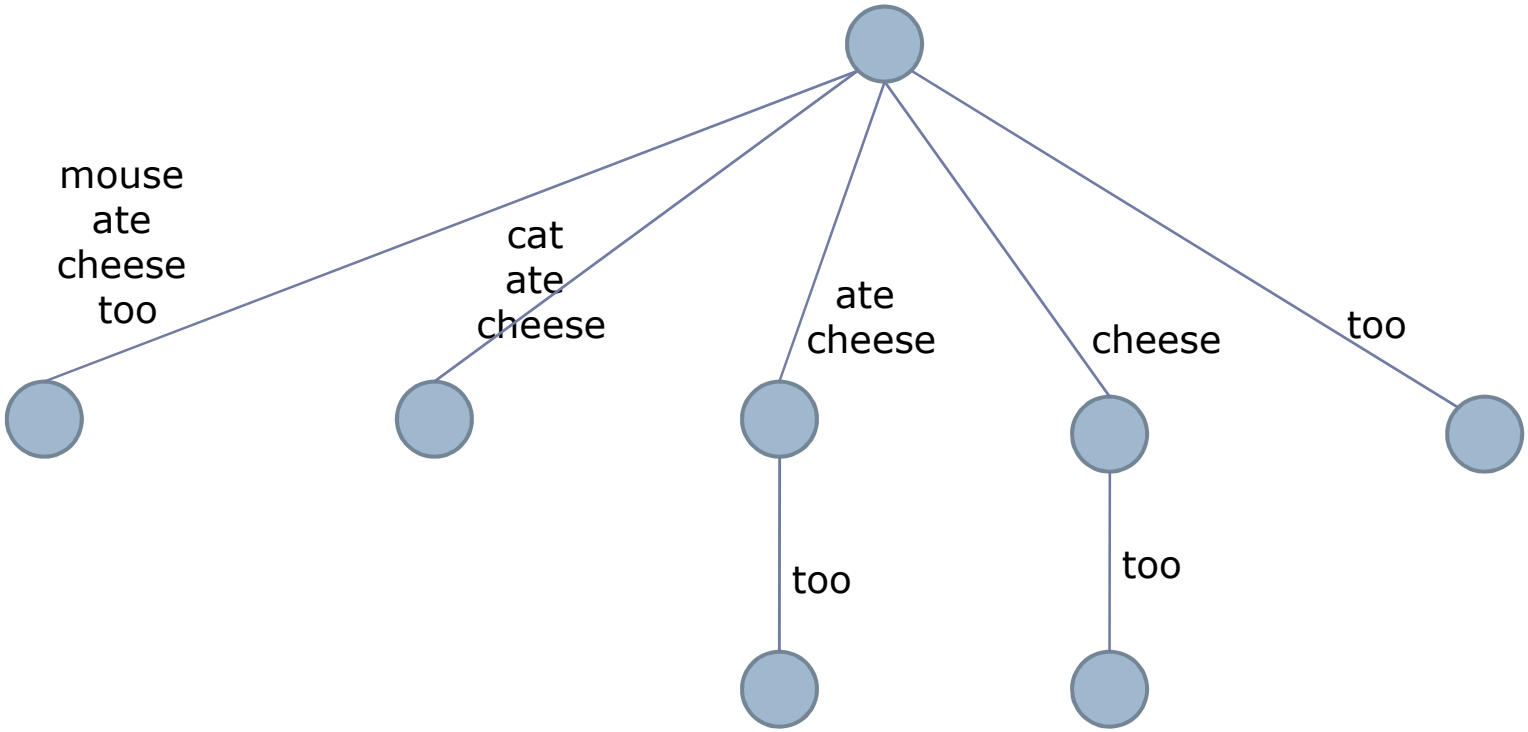
1. cat ate cheese
2. mouse ate cheese too
3. cat ate mouse too



Identifying Base Clusters

Growing our very own Suffix Tree!

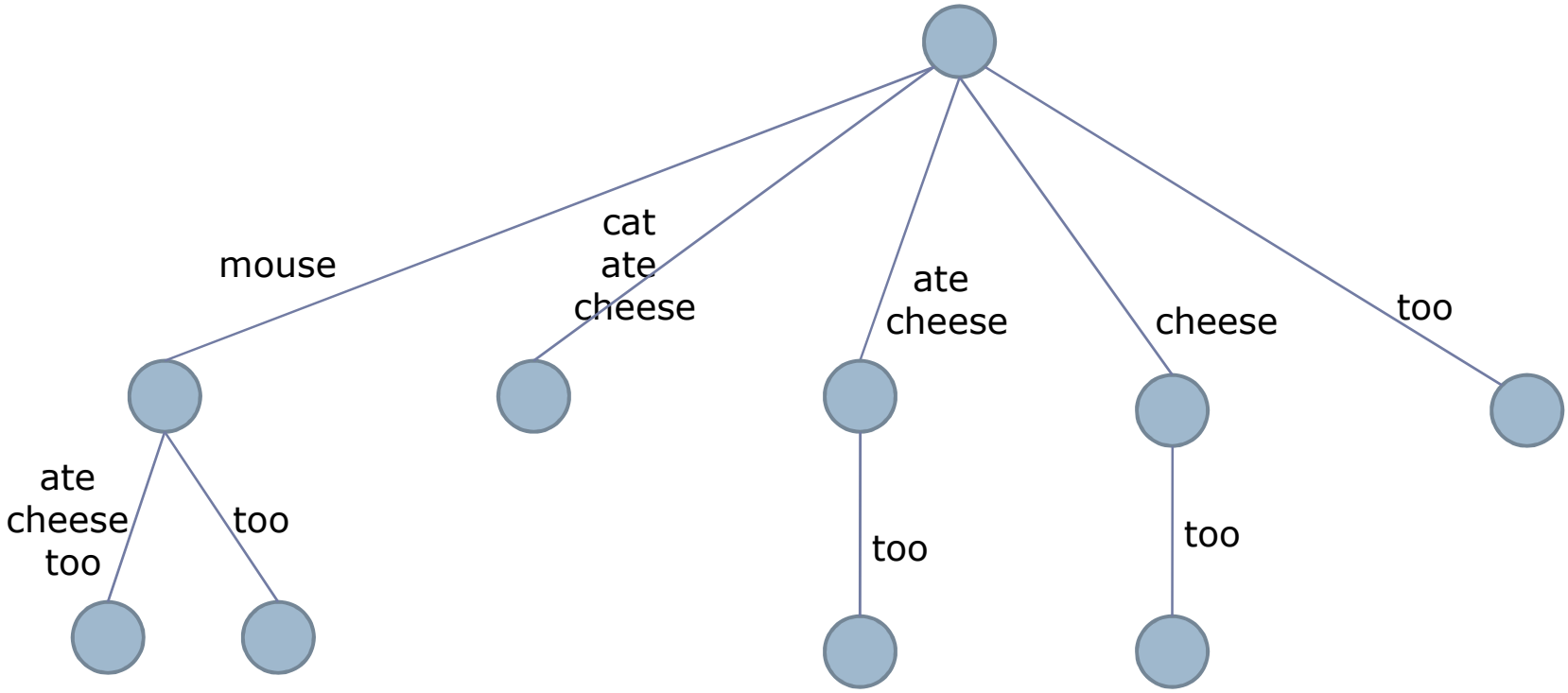
1. **cat ate cheese**
2. **mouse ate cheese too**
3. cat ate mouse **too**



Identifying Base Clusters

Growing our very own Suffix Tree!

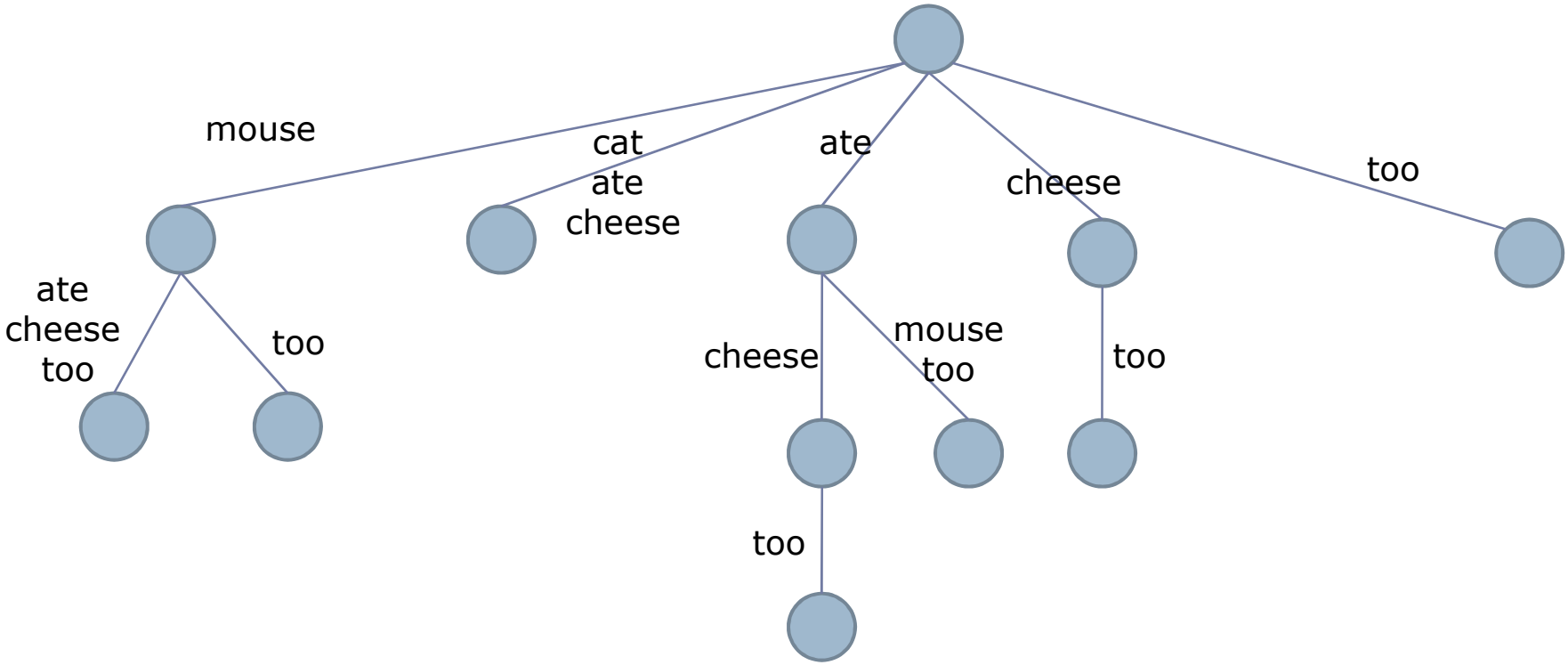
1. **cat ate cheese**
2. **mouse ate cheese too**
3. cat ate **mouse too**



Identifying Base Clusters

Growing our very own Suffix Tree!

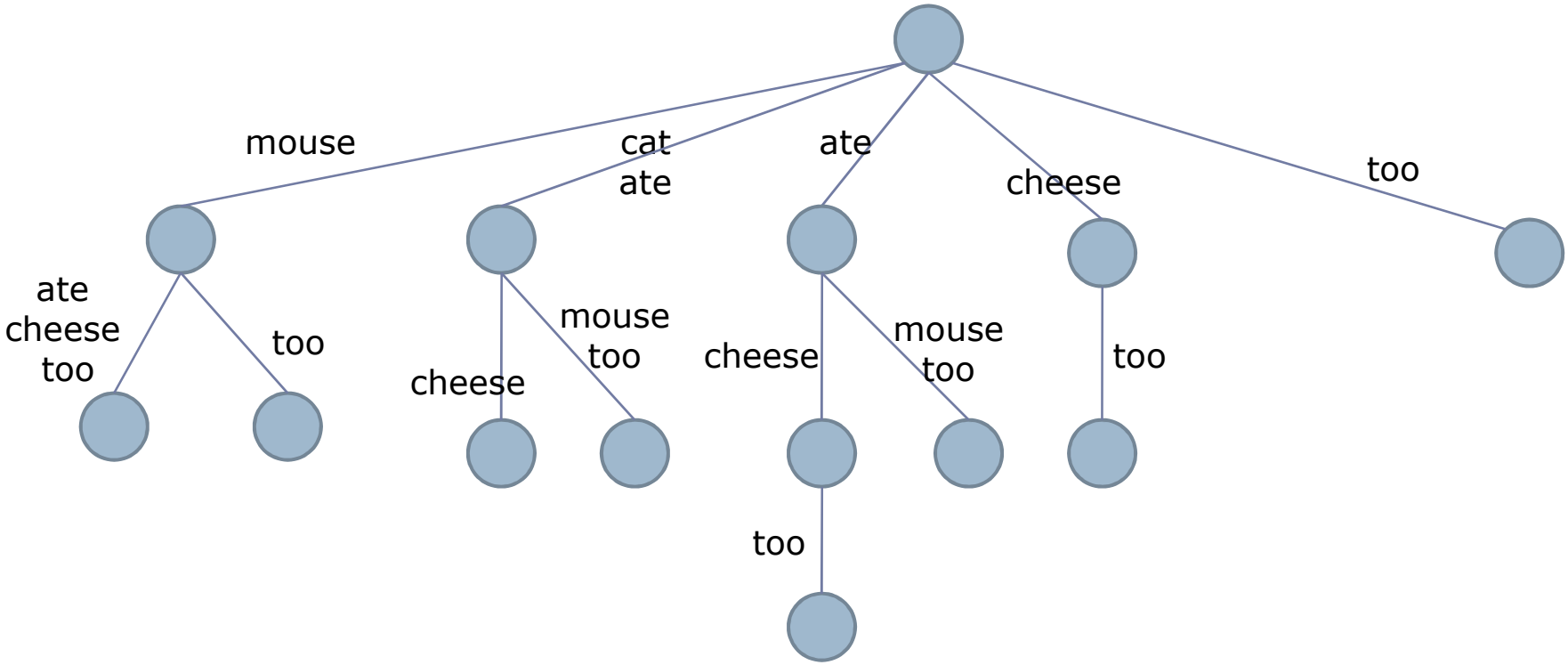
1. **cat ate cheese**
2. **mouse ate cheese too**
3. cat **ate mouse too**



Identifying Base Clusters

Growing our very own Suffix Tree!

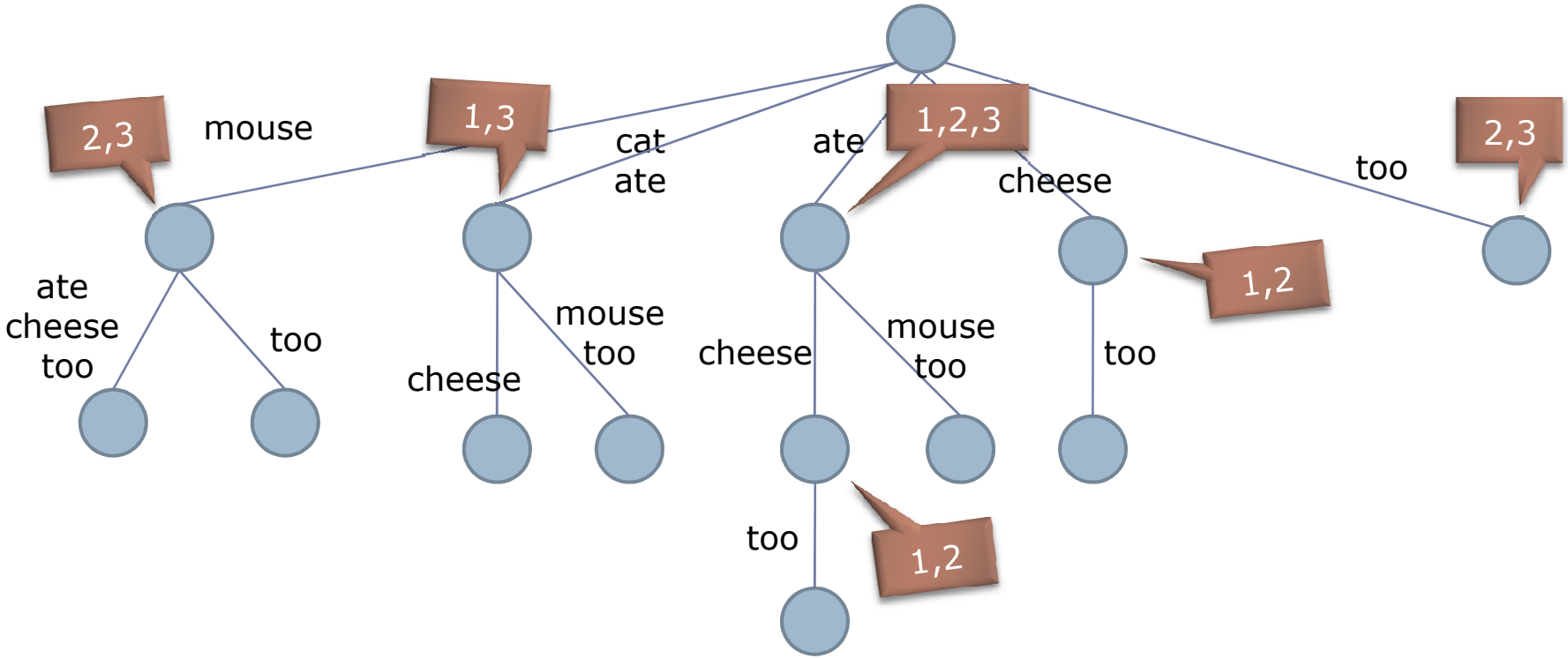
1. cat ate cheese
2. mouse ate cheese too
3. cat ate mouse too



Identifying Base Clusters

Growing our very own Suffix Tree!

1. cat ate cheese
2. mouse ate cheese too
3. cat ate mouse too



Combining Base Clusters

Scoring Clusters

Number of Documents in Cluster

Number of Words in the Cluster

$$s(B) = |B| \cdot f(|P|)$$

Function to penalize single words. Linear for 2-6 words, constant above.

Example:

$$s(ate\ cheese) = 2 \cdot f(2)$$

Combining Base Clusters

Finding similarities

Clusters are similar iff

Documents which are in both clusters

$$\frac{|B_m \cap B_n|}{|B_m|} > 0.5 \quad \text{and} \quad \frac{|B_m \cap B_n|}{|B_n|} > 0.5$$

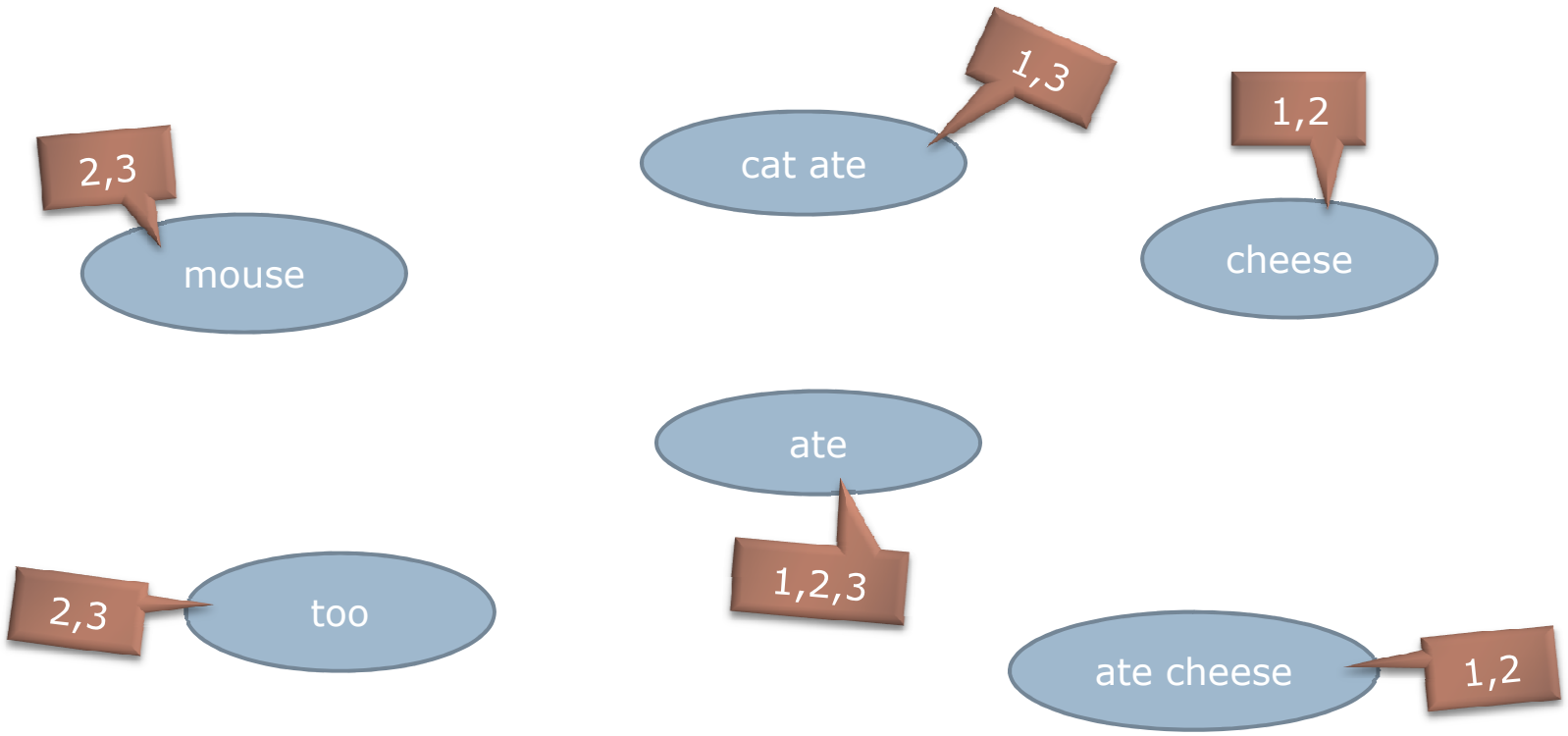
Documents in Cluster m

Documents in Cluster n

Compare new clusters only with the top k scored clusters

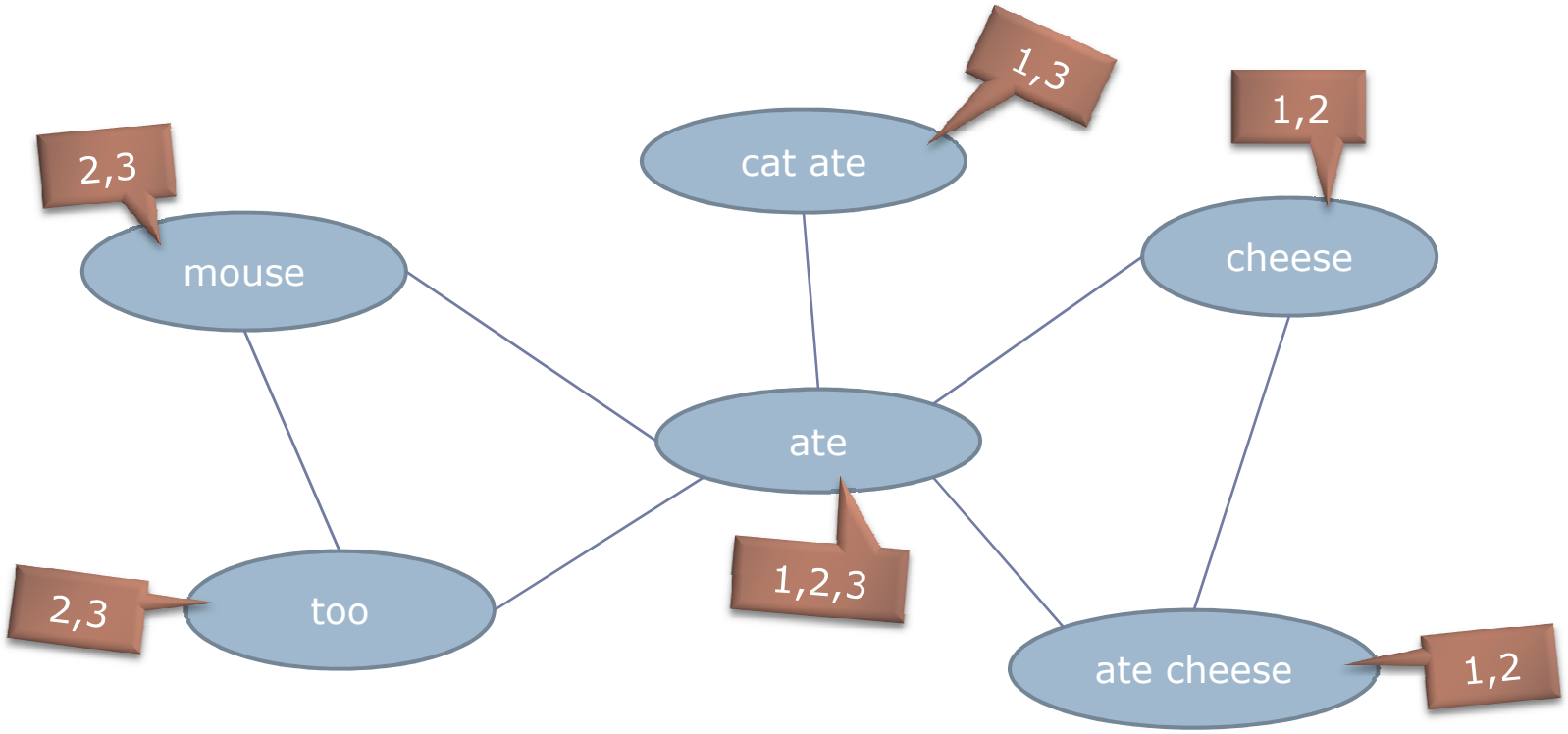
Combining Base Clusters

Base Cluster Graph



Combining Base Clusters

Base Cluster Graph



Experiment And Results

For

1. Effectiveness for Information Retrieval
2. Snippets versus Whole Documents Clustering
3. Execution Time

Evaluation Details

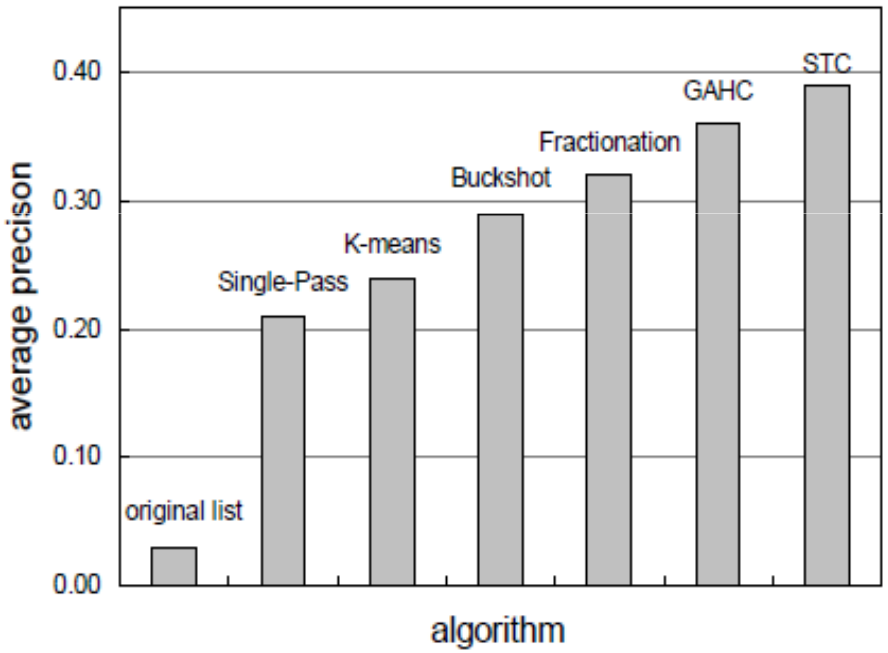
- Comparison with original rank and other clustering algorithms.
- 10 search queries were defined.
- MetaCrawler search engine was used to get test data, using defined queries.
- Top 200 snippets for each of the queries were collected. Also the original document for each snippet was downloaded from the web.
- Each document was manually checked for relevance.
- On average there were 40 relevant documents per query.

Information Retrieval Efficiency Experiment

- Number of clusters produced by each algorithm is a fixed constant [10 in this experiment]
- Similar parameter settings were used, where ever relevant [Eg. Minimum cluster size], which were optimized using a separate dataset.
- These are to allow fair comparison between algorithms.
- Each algorithm tend to create clusters of varying size, and this could artificially influence the comparison between them.
- So only a constant number of documents [10%] were considered, starting from the top cluster and moving down.

Information Retrieval Efficiency Experiment

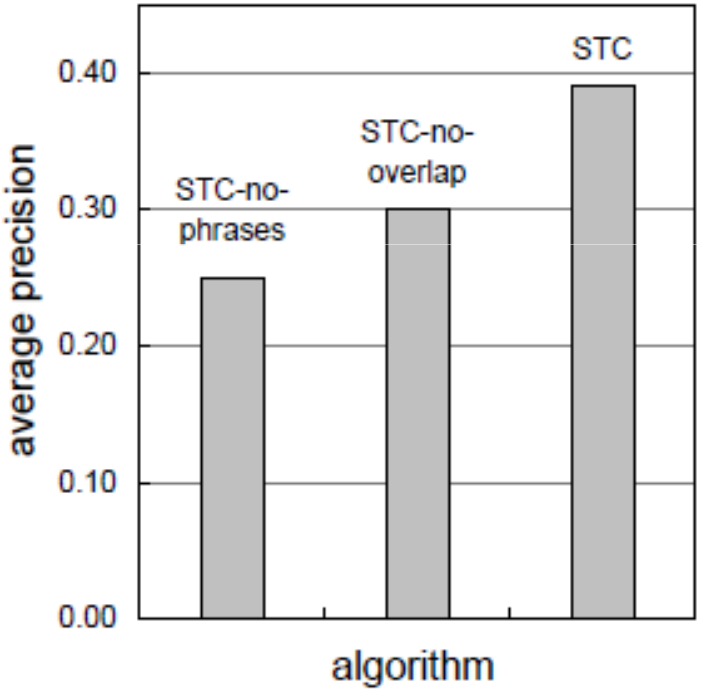
Graph and Some Statistics



- STC scores the highest.
- Reason – Phrases as attribute and Overlapping clusters.
- Average - 2.1 clusters per document.
- 72% of documents were placed in more than one cluster
- 55% of base clusters were based on phrases containing more than one word.

Information Retrieval Efficiency Experiment

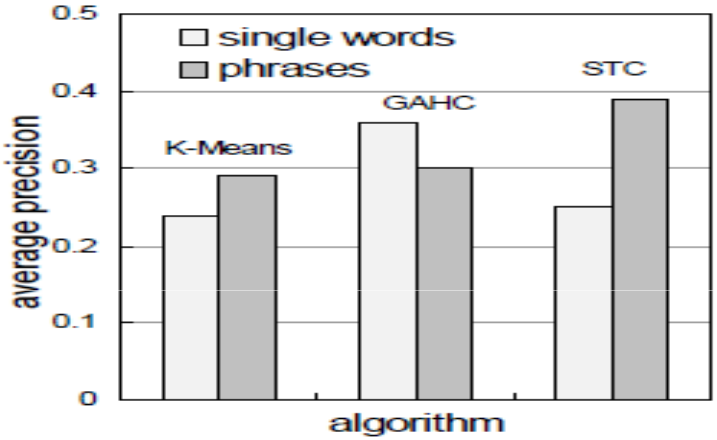
Impact of Phrases and Overlapping Clusters on STC's Performance

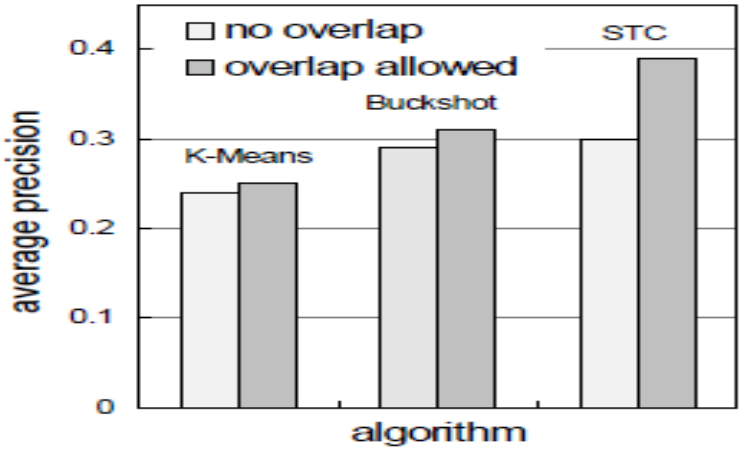


- Remove documents falling in more than one cluster.
- Restrict phrases to one word.
- Performance falls drastically.
- Phrases are basis for identifying cohesive clusters.
- Overlap allows document to feature in all relevant clusters.

Information Retrieval Efficiency Experiment

Can Multi Word Phrase and Overlap Improve the Performance of Other Algorithms?

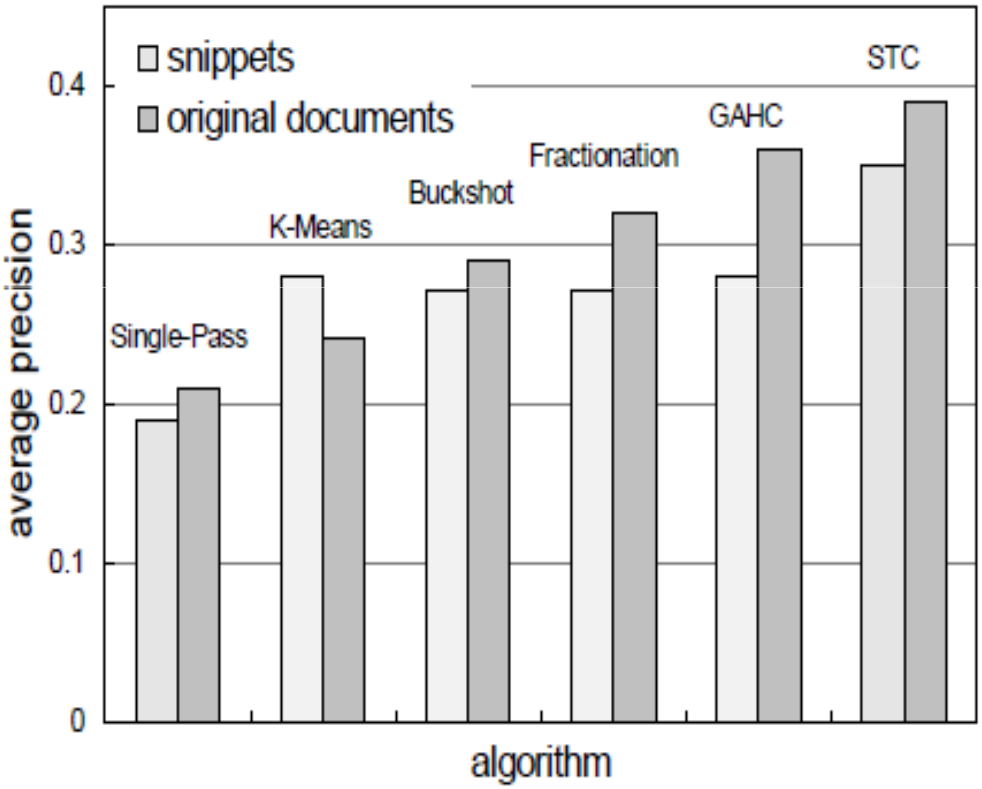




- Either positive or negative impact on vector based algorithms.
- Impact on performance are quit small.
- Degree of cluster overlapping differs.
- Relevant documents appearing in multiple clusters increases the density of relevant documents.
- Irrelevant documents appearing in multiple clusters hurts cluster quality.

	K-Means	Buckshot	STC
Avg. num of clusters: ***Relevant*** document.	1.40	1.40	2.60
Avg. num of clusters: ***Irrelevant*** document	1.55	1.35	1.90
Ratio of the above	0.90	1.04	1.37

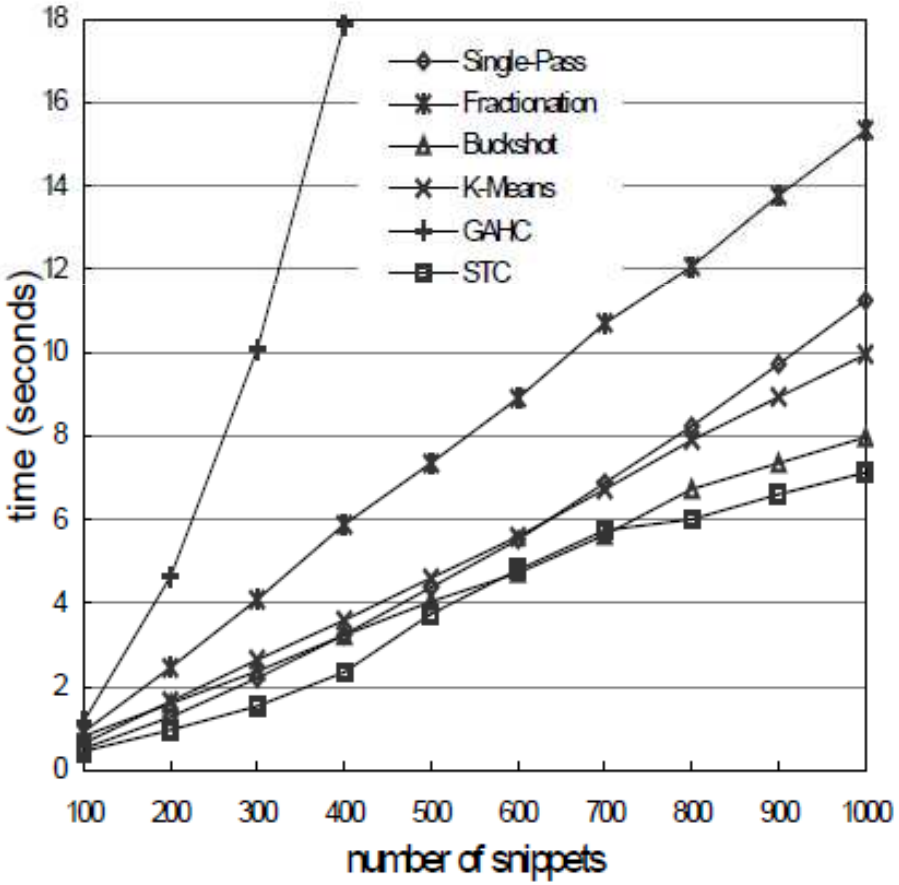
Snippets versus Whole Document



- Web documents – 760 word on average [220 after eliminating stoplist words].
- Snippets – 50 words on average [20 after elimination].
- But decrease in performance is relatively small.
- Possible Reason 1– Snippets contain meaningful phrases and summaries the document well.
- Possible Reason 2 – It omits "noise" contained in the document.

Execution Time

For clustering snippets collection



- Only linear time algorithms is suitable for true online interaction.
- STC is the fastest linear time algorithm in the list.
- STC is even faster since it is incremental in nature.
- STC being incremental makes it utilize the ideal time while waiting for new search results.
- It also enables the system to instantaneously display the results when interrupted by an impatient user.

Live Demo

So how does this work in practice then?

In class we presented an implementation of Suffix Tree Clustering and other algorithms. You can find the open source project here:

http://search.carrot2.org

Under “Cluster with” you can select STC. On the results screen, try clicking on “Visualization”.