ScuDo

Scuola di Dottorato ~ Doctoral School

WHAT YOU ARE, TAKES YOU FAR

Data Mining and Indexing Big Multimedia Data

Mohammad Reza Kavoosifar

Supervisor:

Prof. Elena Baralis

Introduction

- The objectives of my research are:
  1. study and develop a highly scalable indexing scheme for multimodal data
     - The aim of this research is to enable effective content-based multimedia search and retrieval
  2. study and analyze the collaborations between the authors of scientific papers
     - The aim of this research is to understand how authors have collaborated with each other on specific research topics and to what extent their collaborations have been fruitful

- The first research activity has been performed in the context of the TrecVid Hyperlinking task

DIGITAL VIDEO RETRIEVAL at NIST

TRECVID – Hyperlinking task

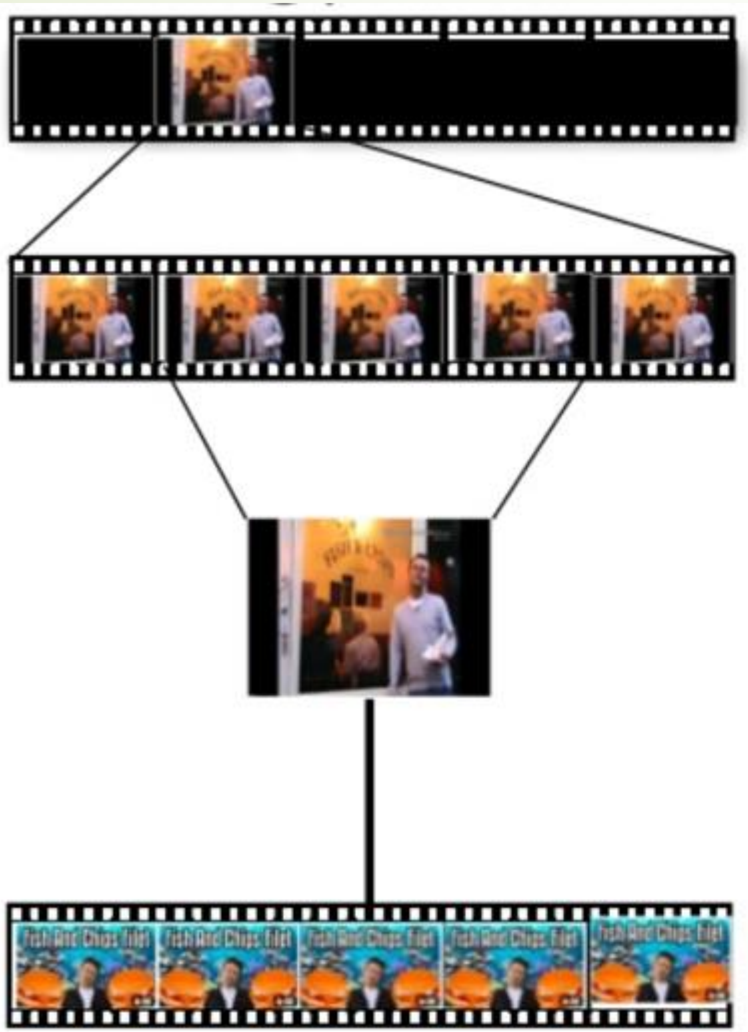
- The goal in video hyperlinking is to suggest relevant video segments based on the multimodal contents of the video segment that a user is currently watching
  - they expect to be provided hyperlinks to related video content within a given archive or collection.
  - There is ambiguity about what the user expectations are regarding these links
    - as well as little information about what is considered relevant to the user in the video segment.
- In this task, one of the main challenges is the uncertainty regarding what criteria are to be followed to generate the links.
  - The task input is a query consisting of an anchor video segment.
  - The task goal is to produce a ranked list of relevant segments with respect to the querying anchor

TRECVID – Hyperlinking task

- **Example:** Consider a video on tourism in London:
  - A video segment (an anchor) on a Fish & Chips restaurant could be linked to a cooking program describing a recipe for Fish & Chips
  - A video segment (an anchor) on the London Parliament could be linked to video segments about England's Queen

https://www-nlpir.nist.gov/projects/tv2016/tv2016.html

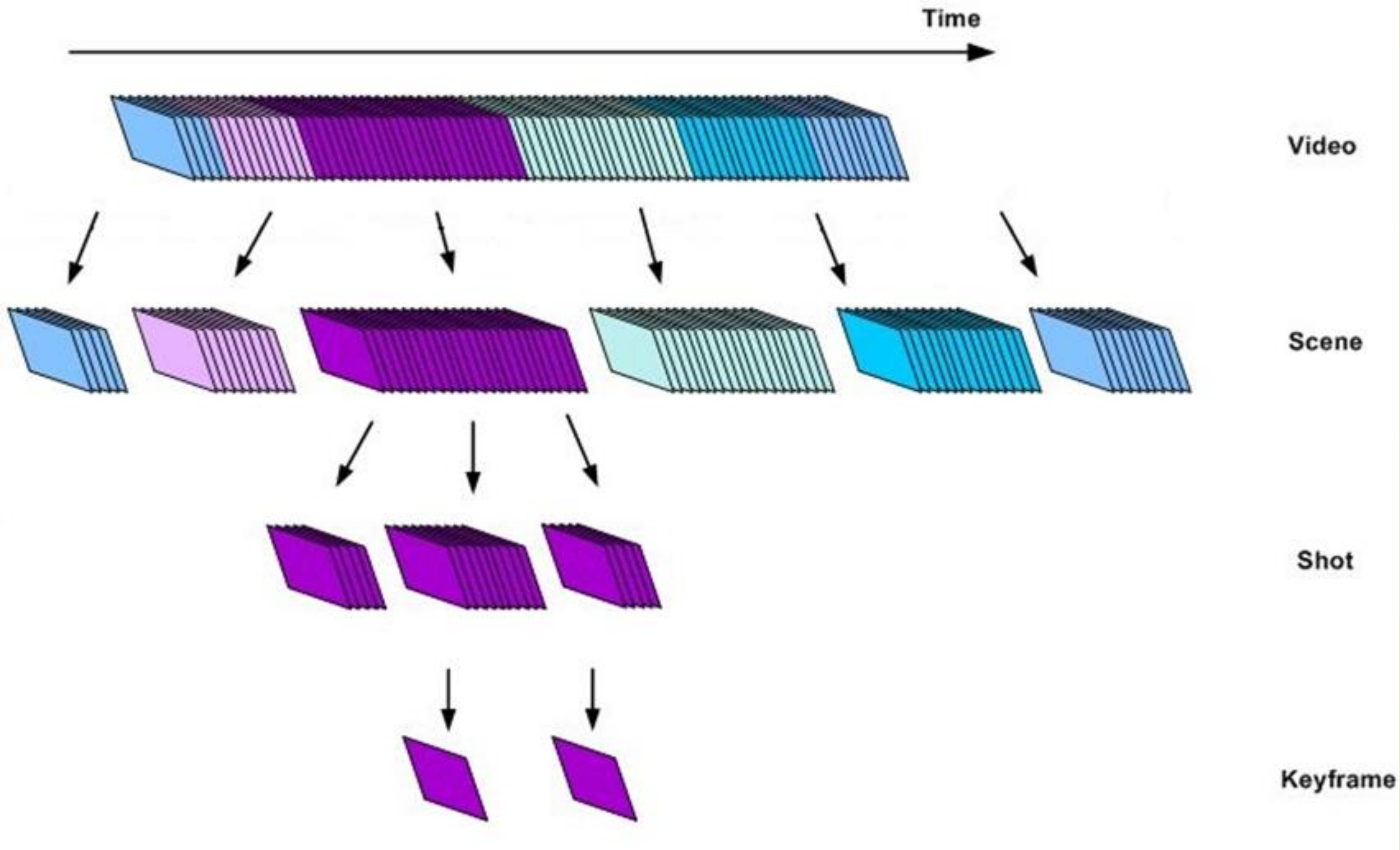
TRECVID – Hyperlinking task



- **Video** (e.g., 2 hours)
- **Video clip** (e.g., 10 min)
- **Anchor**: segment (unconstraint) for which a user requests a link (e.g., 1 min) “I want to know more about this”
- **Hyperlink**
- **Target**: relevant segment for given anchor

https://www.slideshare.net/mariaeskevich/video-hyperlinking-lnk-task-at-trecvid-2016

Video content structuring



http://www.scholarpedia.org/article/video_content_structuring

TRECVID Dataset

- The data of the TRECVID competition are provided by **Blip.tv**
- The dataset consists of **14,838 videos** for a total of **3,288 hours**
- The videos present a variety of topics from computer science tutorials and sightseeing guides to homemade song covers.
- Videos are characterized by
  - **Metadata** (title, short program descriptions,…)
  - **Automatic speech recognition (ASR) transcripts** (LIUM and LIMSI)
  - **Visual concepts**
    - extracted using the Caffe framework with the BVLC GoogLeNet model
      - trained to classify images into 1000 different ImageNet categories.
  - Shots and Keyframes
- Training set
  - 94 query anchors with a set of ground-truth relevant related segments are provided
- Test set
  - 25 query anchors

TRECVID Dataset: Visual concepts

- Visual concepts are the concepts which are being detected in a keyframe by exploiting an image processing tool.

- Detected concepts:
  - golf ball
  - croquet ball
  - racket
  - Ballplayer
  - baseball player
- Skipped concepts due to low confidence score :
  - ~~Gar~~
  - ~~Garfish~~
  - ~~Billfish~~

Features considered for the system

- We proposed a system based on different combinations of both textual and visual features.
- We used
  - **LIMSI** Automatic speech recognition (ASR) transcripts
  - **Visual concepts**
  - **Metadata**
- We also considered extra features to identify the most relevant terms and concepts in each query
  - Named-entity recognition (NER)
  - Concept mapping technique
    - Using Wordnet

Named Entity Recognition (NER)

- Named Entity Recognition labels sequences of words in a text which are the names of things, such as person and company names, ...
- I used Stanford Named Entity Recognizer (NER)
  - From Stanford university
- Stanford NER is also known as CRFClassifier.
  - It provides a general implementation of (arbitrary order) linear chain Conditional Random Field (CRF) sequence models.
- NER is used to **assign a higher relevance** to those words that are entities.
  - The basic idea is that the segments containing the entities appearing in the anchor are potentially more interesting.
  - NER never used alone as a monomodal query and it is always combined with another feature like LIMSI transcripts.
- For example:
  - ✓ In this video PGA Tour player Heath Slocum speaks about his experience with instruction.

Concept mapping technique

- Concept mapping technique is used to find the most relevant concepts inside the query.
  - is based on **WordNet**
  - The mapping is done by using the words appearing in meta-data of the video and the concepts list of the segment.
  - In order to enrich the words list, we applied WordNet using the synonyms and hypernyms of the words.
    - A hypernym is a word with a broad meaning constituting a category into which words with more specific meanings fall; a super-ordinate.
    - For example, **color is a hypernym of red**

Concept mapping technique

- The concept mapping technique **tries to increase the relevance** of the visual concepts of the considered anchor that are related to the content of the whole video.
  - each visual concept of the anchor is compared with the words appearing in the metadata of the video containing the anchor.
  - If the visual concept, or its synonymous based on Wordnet, appears in the metadata of the video then the weight of that visual concept is increased
- For example:
  - metadata title is: "Top 100 golf tips for kids"
  - Visual concepts are: "digital clock, golf ball"
  - ✓ golf ball is selected since golf is matching

System overview

- The proposed system has 3 distinct stages

1. **Data segmentation**
   - We considered 120-seconds Fixed-segmentation
   - We also applied **data cleaning**
2. **Indexing**
   - Apache Solr was used to index the data
3. **Query formulation and segment retrieval**
   - Selecting the most relevant segments

1) Data segmentation

- The goal in this step is to split the videos in segments.
- We used a **120 sec Fixed-segmentation**
- Based on our previous experiments, **Shot-segmentation is not a good choice** to investigate
  - Because the videos are a collection of semi-professional user-generated videos where they are not edited and for most of them, people filmed themselves.
- Also Fixed-segmentation seem to provide better coverage and more choice than the lower length segmentation.
  - the 120 seconds is the upper bound for an anchor in the Hyperlinking task
  - the minimum length is 10 seconds.

Data cleaning

- All the textual data associated with the segments have been preprocessed to remove irrelevant words.
- We used:
  - punctuation removal tool
  - Stopword elimination tool

- Stopword elimination tool
  - The words occurring in the textual data are compared with those contained in a dictionary of conjunctions, articles, prepositions, abbreviations etc. and matching words are removed.
  - We used 665 different English stop-words

- We also applied stemming; however, it is integrated in the Indexing part.

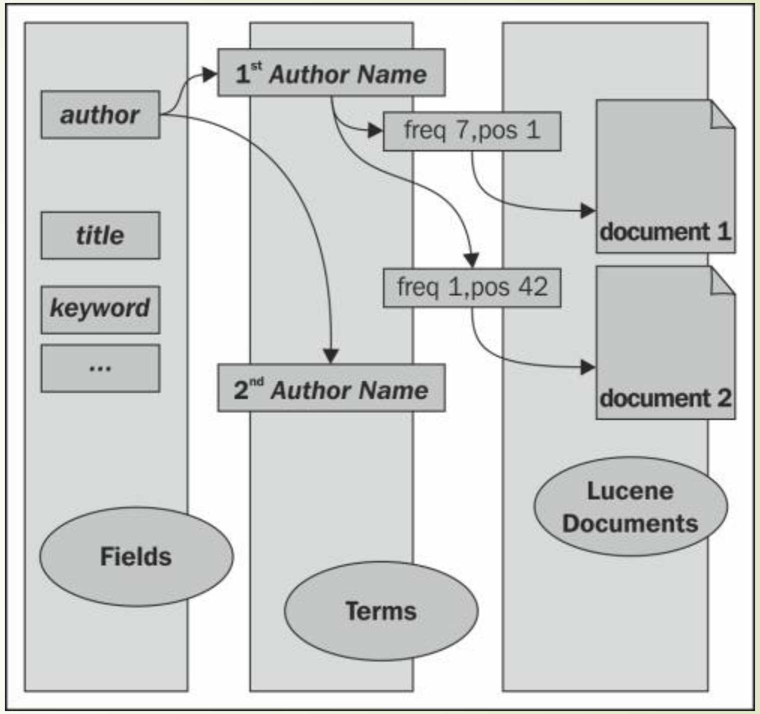
2) Indexing

- Indexes created for the video segments based on one of the following features:
  1. the **LIMSI** transcripts of the segments
  2. the **visual concepts** of the segments
  3. the **metadata** of the full videos

- **Apache Solr** has been used to index the textual and visual features associated with each segment.

Apache Solr

- The indexing structure implemented by Solr is known as **inverted index**.
  - An inverted index stores, for each term, the list of documents where the term is present.
  - This makes term-based queries very efficient



- Stemming:
  - During the creation of indexes, a stemming algorithm is applied on the document.
  - We exploited **SnowballPorter** algorithm.

Apache Solr: Relevance score

- Solr uses a formula called **Practical Scoring Function** to calculate relevance.

$$\text{score}(q,d) = \text{coord}(q,d) \cdot \text{queryNorm}(q) \cdot \sum_{t \text{ in } q} \left(\text{tf}(t \text{ in } d) \cdot \text{idf}(t)^2 \cdot \text{t.getBoost}() \cdot \text{norm}(t,d) \right)$$

- The standard similarity algorithm used in Solr is term frequency/inverse document frequency, or **TF/IDF**.
- **Query Boosting**:
  - Not all terms are equally important in a query
  - A query boost is a factor that Solr considers when computing a score
    - **a higher boost value returns a higher score.**
- Query normalization:
    - **queryNorm**(q) is a normalizing factor **used to make scores between queries (or even different indexes) comparable**.

3) Query formulation and segment retrieval

- The goal in this stage is to generate an optimal query text to be used for the segment retrieval on Solr indexes.
- The proposed approach is designed to build an enriched query text from the available features:
  - the video query segment (anchor) is converted into a textual query string by considering a combinations f the following features:
    1. all the textual information associated with the anchor
       - LIMSI transcripts
       - visual concepts
    2. the metadata of the full video containing the anchor
       - Title, Description and tags
    3. additional text obtained by:
       - Named Entity Recognition (NER)
       - Concept Mapping technique
- Finally, the enriched query is used to query the Apache Solr indexes
  - hence identifying related segments, ranked by relevance.

Mono-modal Query formulation

- In the proposed system, we considered four different mono-modal queries:

1. LIMSI-based query + Named-Entity Recognition
2. Visual-concept-based query + concept mapping technique
3. Metadata-based query for segment selection
4. Metadata-based query for video selection

1) LIMSI-based query + NER

- A textual query is built by
  1. considering the words appearing in the **LIMSI** transcript of the anchor
  2. Named-Entity Recognition (NER) is applied on the anchor LIMSI transcripts
     - to extract relevant names of entities
     - give them higher relevance in the query

- For example, if the LIMSI text is:

  "Handmade portraits: Staceyrebecca",

  - ✓ the query would be:

    "Handmade" (W1.0) OR "portraits" (W1.0) OR "Staceyrebecca" (W1.6)

2) Visual-concept-based query + concept mapping technique

- For each video anchor, a textual query is built by considering the "names" of visual concepts appearing in the anchor.
- The visual concepts with a **score greater than 0.3**, as provided by the GoogleNet model, are selected
- For example:
  - metadata text is: “Top 100 golf tips for kids”
  - Visual concepts are: “digital clock, golf ball”
  - ✓ " digital clock” (W1.0) OR “golf ball " (W1.6)

Metadata-based query

- Metadata are associated to the full video
  - If the query is executed on a metadata index, only full videos can be selected, with all their corresponding segments.

3. **For segment selection**
   - metadata queries are executed on the **LIMSI transcript index**
     - Because transcripts are specific for each segment
   - Named-Entity Recognition (NER) is applied
     - to extract relevant entities and give them higher relevance in the query

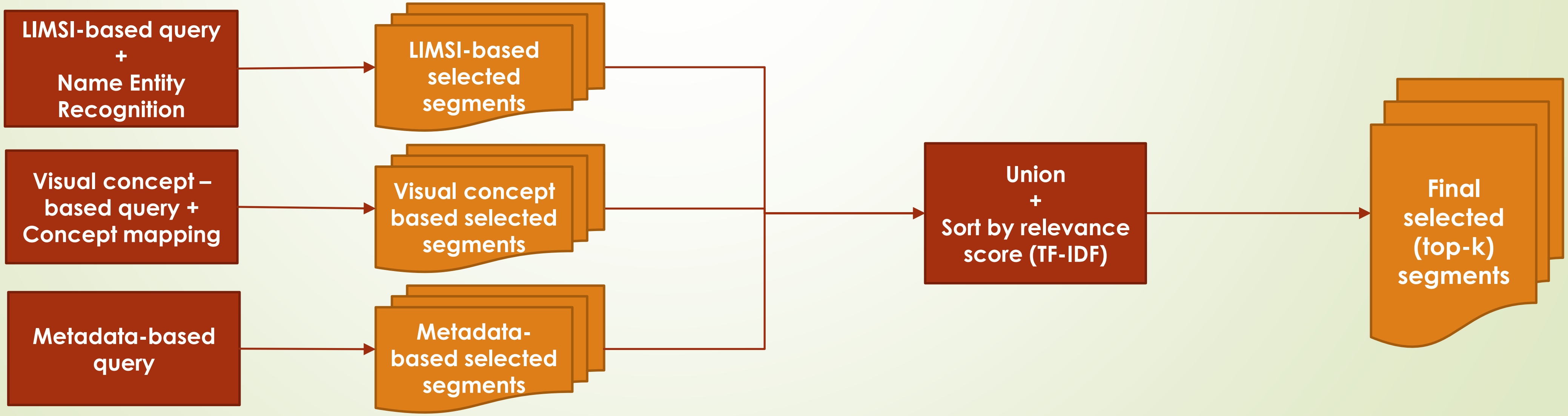
4. **For video selection**
   - it is executed on the metadata index
     - Returning videos and not segments
   - the results of such query cannot be used directly to propose the resulting segments
     - this query helps in **filtering a pre-selection of videos** among which related segments are highly likely to be found

Hyperlinking algorithms

- In the proposed system, we combined multiple mono-modal queries into a globally multi-modal system:
  - The novelty of the algorithms is the way they use the provided features
  1. Automatic Feature selection (AFS)
  2. Metadata based approach
  3. Pipeline approach
- We considered also a monomodal algorithm
  - Based on LIMSI transcript with Named-entity recognition (NER)
  - it is considered because it is a core part embedded in the other proposed combinations
    - its separate evaluation is a noteworthy addition for the experimental comparison to identify its specific contribution to the overall results.

Hyperlinking algorithms

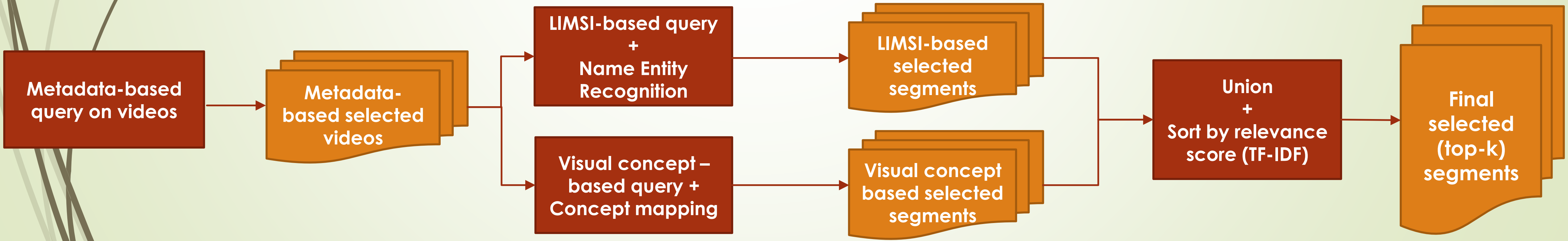
1. **Automatic Feature selection (AFS)**
   - Features: Metadata, LIMSI, Visual concepts
     - Also Named-entity recognition (NER) and Concept mapping technique
   - For each anchor:
     - Select one set of relevant segments for each feature by considering one feature at a time
     - Consider the union of the selected segments and select the subset of segments with the highest relevance score
   - We used the TF-IDF score to identify the relevance score of each selected segment



Hyperlinking algorithms

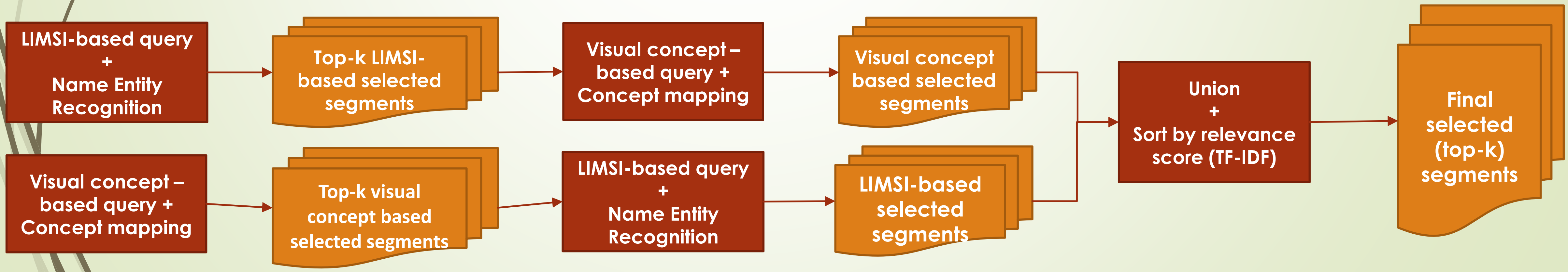
2. Metadata based approach

- Features: Metadata, LIMSI, Visual concepts
  - Also Named-entity recognition (NER) and Concept mapping technique
- For each anchor:
  - Select relevant videos by using metadata for querying the video collection
  - Select the most relevant segments from the selected videos by using LIMSI and visual concepts



Hyperlinking algorithms

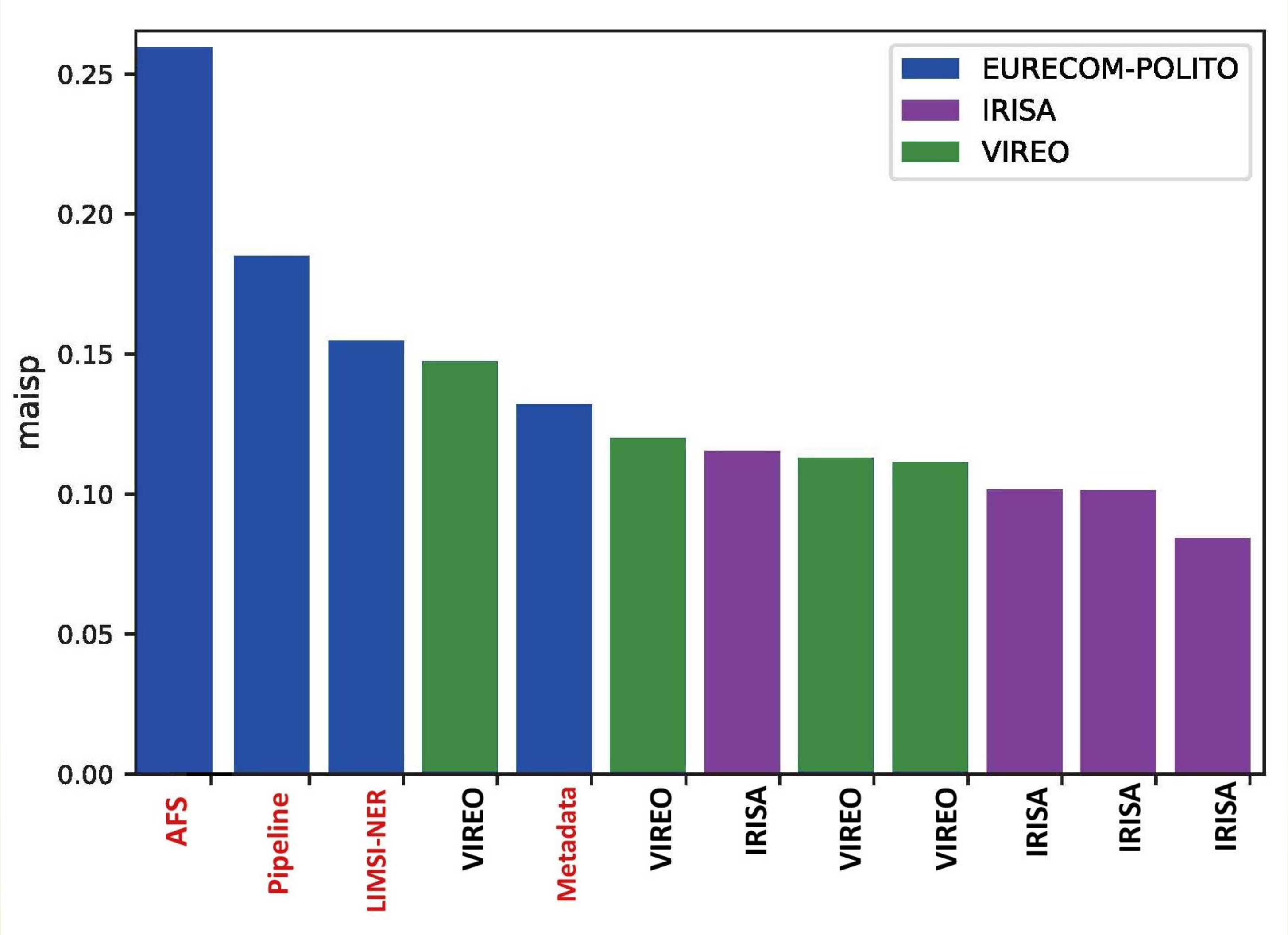
3. **Pipeline approach**
   - Features: LIMSI, Visual concepts
     - Also Named-entity recognition (NER) and Concept mapping technique
   - For each anchor:
     - **Step 1-1:** Select relevant videos by using LIMSI for querying the video collection
     - **Step 1-2:** Select the most relevant segments from the selected videos by using visual concepts
     - **Step 2:** Repeat the step 1 by switching the roles of LIMSI and visual concepts
     - **Step 3:** Consider the union of the selected segments and select the subset of segments with the highest relevance score
   - Here is the schema for the first step of this algorithm:



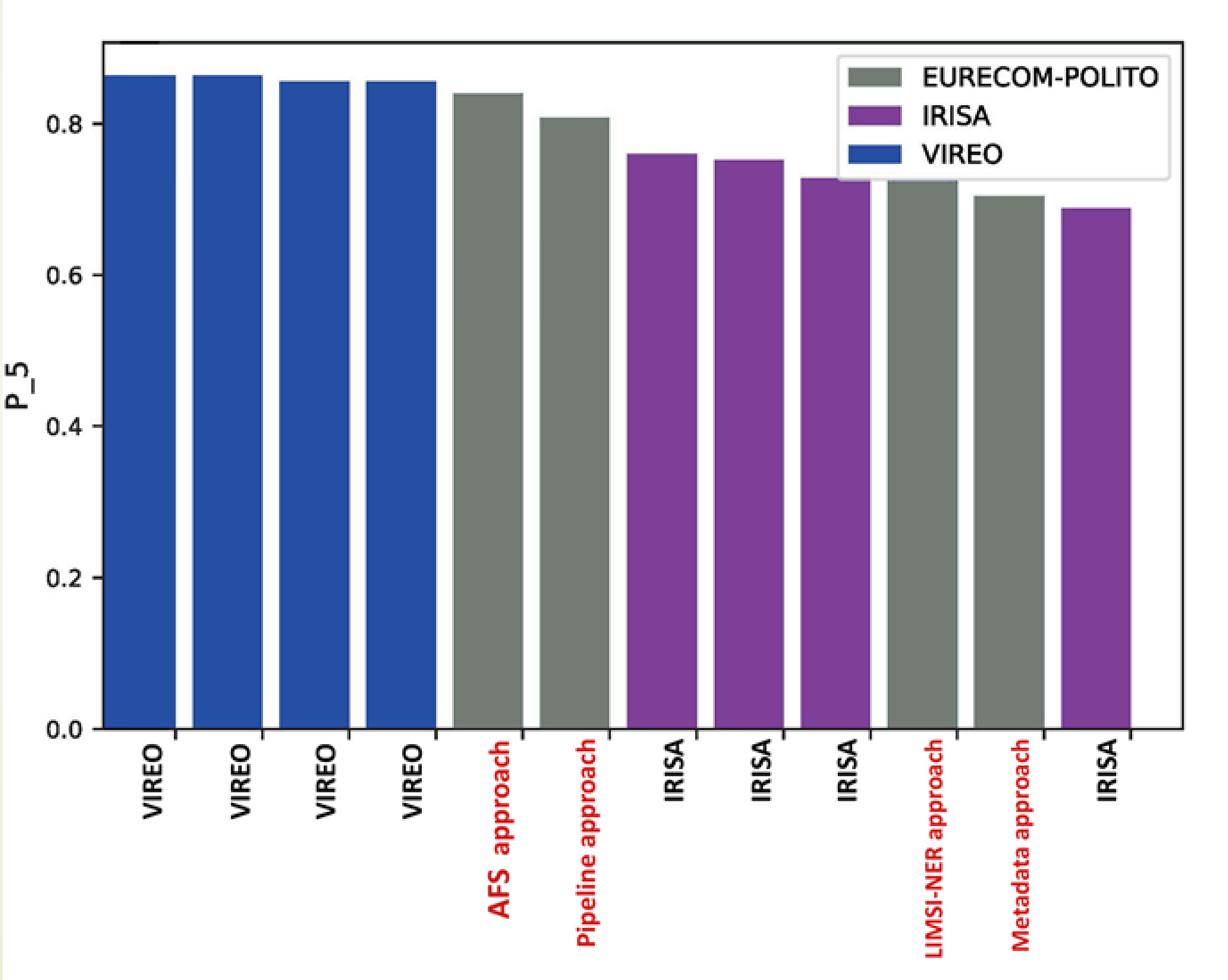
Evaluation metrics

- Results have been evaluated according to the following metrics:
  - Precision at rank 5 (P@5)
    - i.e., the number of true positives in the top 5 selected segments.
  - Precision at rank 10 (P@10).
  - Mean Average Precision (MAP)
    - considers true positives all segments overlapping with a segment that was considered relevant in the ground truth
  - Mean Average interpolated Segment Precision (**MAiSP**)
    - adapted from MAP
    - includes rewards/penalties for segmentation accuracy

Results at TrecVid (MAiSP)



Results at TrecVid (Precision @5)



Analysis on the impact of parameters

- An analysis is being done on the impact of parameters for the developed algorithms
  - To improve the performance of algorithms
- The standard configuration considered for all approaches when analyzing the pre-evaluation results:

1. Top K-filter: **1000**
2. Stemming algorithm: **SnowballPorter**
3. Visual concepts filter threshold: **0.3**
4. Query boost value: **1.6**
5. NER classifier: **Multi Classifier**
6. WordNet similarity algorithms: **Lin**
   - Lin algorithm threshold: **0.7**

Analysis on the impact of parameters

Algorithm	AFS	Metadata	Pipeline	LIMSI-NER
SnowballPorter	**0.289**	**0.227**	**0.221**	**0.212**
PorterStem	0.278	0.215	0.205	0.198
Hunspell	0.224	0.187	0.178	0.153
KStem	0.219	0.181	0.173	0.145

P@10 results for stemming algorithms in Solr

Threshold	AFS	Metadata	Pipeline
0.2	0.256	0.219	0.213
0.3	**0.289**	**0.227**	**0.221**
0.5	0.243	0.211	0.207
0.7	0.231	0.204	0.198

P@10 results for filter threshold of visual concepts

Boost value	AFS	Metadata	Pipeline	LIMSI-NER
1.2	0.268	0.211	0.202	0.197
1.3	0.268	0.211	0.202	0.197
1.4	0.273	0.215	0.208	0.202
1.5	0.281	0.221	0.215	0.208
1.6	**0.289**	**0.227**	**0.221**	**0.212**
1.7	0.283	0.223	0.217	0.209
1.8	0.280	0.219	0.214	0.206

P@10 results for query boost value

Classifier	AFS	Metadata	Pipeline	LIMSI-NER
No Classifier	0.197	0.164	0.152	0.136
Single Classifier	0.271	0.210	0.207	0.193
Multi Classifier	**0.289**	**0.227**	**0.221**	**0.212**

P@10 results for NER classifiers

Visualization

LIMSI 2016 | LIUM 2012 | Visual concepts (Filter 0.3) | Visual concepts (Filter 0.5) | MetaData | Pipeline | AFS

Anchor: **anchor_66** Change anchor

Query

video: **vid04712**, Time: **00:02:45 - 00:03:30**

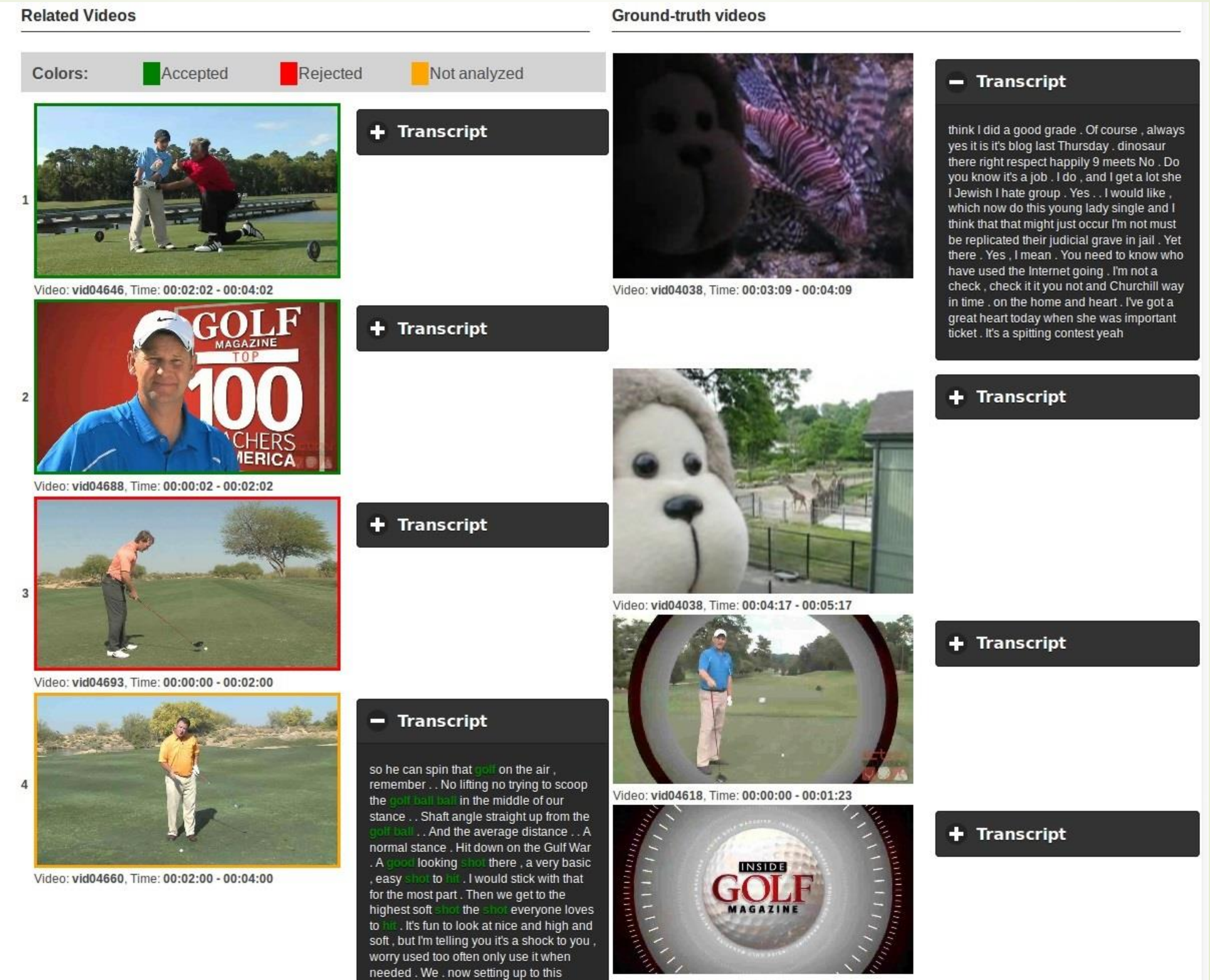
Transcript

there , I do You can see customers walking into the mall he's gonna take a practice swing versus what we're looking for and then he's in a walking in with his right foot place a club behind the ball . move his feet shoulder width apart needs and flexed he's checking now to see how far the violent clubs from him and one of the things customer has done is he's come up with a system that helps him of good rhythm , timing when he takes the club back . He says to himself . Tiger Woods Somalia dirt and Tiger Woods if you want your child to play good golf game , a good grip tightens swing up at the magic teaching DA will give him a real good pre-shot routine and then let had something simple like Tiger Woods and hit the ball as hard as they can and that will make them a great golfer

Query:

custom OR walk OR mall OR practic OR swing OR foot OR place OR club OR ball OR move OR feet OR shoulder OR width OR apart OR flex OR check OR violent OR thing OR system OR help OR good OR rhythm OR time OR take OR "tiger woods"^1.6 OR "somalia"^1.6 OR dirt OR child OR play OR "golf"^1.6 OR game OR grip OR tighten OR magic OR teach OR da OR real OR shot OR routin OR simpl OR hit OR hard OR great OR golfer

Visualization



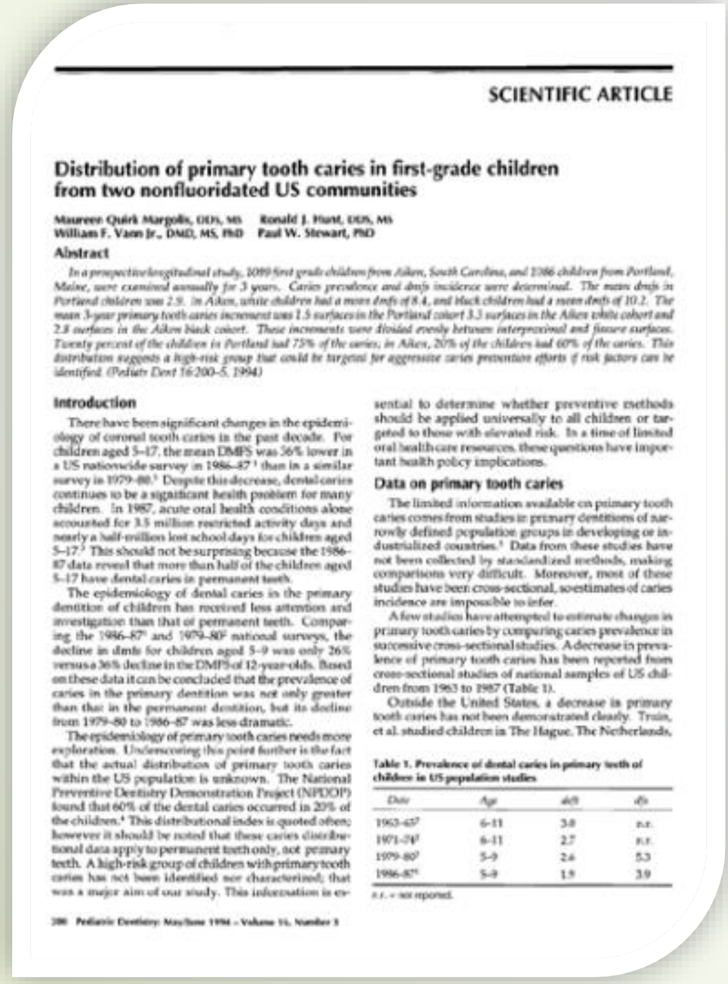
Future work

- An improvement could be applied to the Automatic Feature Selection (AFS) algorithm
  - Using the new data features like OCR
- Dynamic segmentation
    - For example, considering the end of sentences by LIMSI

- Other TRECVID tasks:
  - Ad-hoc Video Search (AVS) task
    - The idea is to promote the development of methods that permit the indexing of concepts in video shots using only data from the Web or archives without the need of additional annotations.
  - Streaming Multimedia Knowledge Base Population (SMKBP)
    - Goal: extract Knowledge elements, about events, actions, ... from a variety of unstructured sources

Discovering cross-topic collaborations among researchers

Discovering cross-topic collaborations among researchers



SCIENTIFIC ARTICLE

Distribution of primary tooth caries in first-grade children from two nonfluoridated US communities

Maureen Quirk Margolis, DDS, MS Ronald J. Hunt, DDS, MS
William F. Vann Jr., DMD, MS, PhD Paul W. Stewart, PhD

Abstract

In a prospective longitudinal study, 1099 first grade children from Aiken, South Carolina, and 1386 children from Portland, Maine, were examined annually for 3 years. Caries prevalence and dmfs incidence were determined. The mean dmfs in Portland children was 2.9. In Aiken, white children had a mean dmfs of 8.4, and black children had a mean dmfs of 10.2. The mean 3-year primary tooth caries increment was 1.5 surfaces in the Portland cohort 3.3 surfaces in the Aiken white cohort and 2.8 surfaces in the Aiken black cohort. These increments were divided evenly between interproximal and fissure surfaces. Twenty percent of the children in Portland had 75% of the caries; in Aiken, 20% of the children had 60% of the caries. This distribution suggests a high-risk group that could be targeted for aggressive caries prevention efforts if risk factors can be identified. (Pediatr Dent 16:200–5, 1994)

Introduction

There have been significant changes in the epidemiology of coronal tooth caries in the past decade. For children aged 5–17, the mean DMFS was 36% lower in a US nationwide survey in 1986–87[1] than in a similar survey in 1979–80.[1] Despite this decrease, dental caries continues to be a significant health problem for many children. In 1987, acute oral health conditions alone accounted for 3.5 million restricted activity days and nearly a half-million lost school days for children aged 5–17.[3] This should not be surprising because the 1986–87 data reveal that more than half of the children aged 5–17 have dental caries in permanent teeth.

The epidemiology of dental caries in the primary dentition of children has received less attention and investigation than that of permanent teeth. Comparing the 1986–87[1] and 1979–80[2] national surveys, the decline in dmfs for children aged 5–9 was only 26% versus a 36% decline in the DMFS of 12-year-olds. Based on these data it can be concluded that the prevalence of caries in the primary dentition was not only greater than that in the permanent dentition, but its decline from 1979–80 to 1986–87 was less dramatic.

The epidemiology of primary tooth caries needs more exploration. Underscoring this point further is the fact that the actual distribution of primary tooth caries within the US population is unknown. The National Preventive Dentistry Demonstration Project (NPDDP) found that 60% of the dental caries occurred in 20% of the children.[4] This distributional index is quoted often; however it should be noted that these caries distributional data apply to permanent teeth only, not primary teeth. A high-risk group of children with primary tooth caries has not been identified nor characterized; that was a major aim of our study. This information is essential to determine whether preventive methods should be applied universally to all children or targeted to those with elevated risk. In a time of limited oral health care resources, these questions have important health policy implications.

Data on primary tooth caries

The limited information available on primary tooth caries comes from studies in primary dentitions of narrowly defined population groups in developing or industrialized countries.[5] Data from these studies have not been collected by standardized methods, making comparisons very difficult. Moreover, most of these studies have been cross-sectional, so estimates of caries incidence are impossible to infer.

A few studies have attempted to estimate changes in primary tooth caries by comparing caries prevalence in successive cross-sectional studies. A decrease in prevalence of primary tooth caries has been reported from cross-sectional studies of national samples of US children from 1963 to 1987 (Table 1).

Outside the United States, a decrease in primary tooth caries has not been demonstrated clearly. Truin, et al. studied children in The Hague, The Netherlands,

Table 1. Prevalence of dental caries in primary teeth of children in US population studies

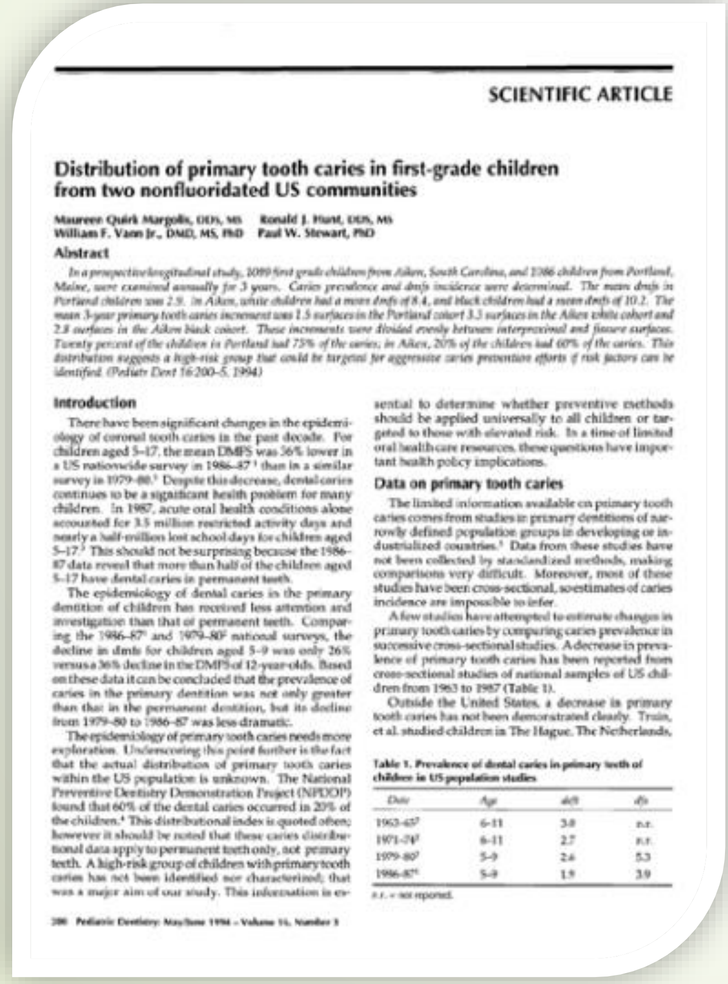
Date	*Age*	*deft*	*dfs*
1963–65[7]	6–11	3.0	n.r.
1971–74[3]	6–11	2.7	n.r.
1979–80[2]	5–9	2.6	5.3
1986–87[1]	5–9	1.9	3.9

n.r. = not reported.

- **How to find the scientific publications of major interest?**

1. Topic-driven searches
2. Author-driven searches
3. All of the above

Discovering cross-topic collaborations among researchers



SCIENTIFIC ARTICLE

Distribution of primary tooth caries in first-grade children from two nonfluoridated US communities

Maureen Quirk Margolis, DDS, MS Ronald J. Hunt, DDS, MS
William F. Vann Jr., DMD, MS, PhD Paul W. Stewart, PhD

Abstract

In a prospective longitudinal study, 1099 first grade children from Aiken, South Carolina, and 1386 children from Portland, Maine, were examined annually for 3 years. Caries prevalence and dmfs incidence were determined. The mean dmfs in Portland children was 2.5. In Aiken, white children had a mean dmfs of 8.4, and black children had a mean dmfs of 10.2. The mean 3-year primary tooth caries increment was 1.5 surfaces in the Portland cohort 3.3 surfaces in the Aiken white cohort and 2.8 surfaces in the Aiken black cohort. These increments were divided evenly between interproximal and fissure surfaces. Twenty percent of the children in Portland had 75% of the caries; in Aiken, 20% of the children had 60% of the caries. This distribution suggests a high-risk group that could be targeted for aggressive caries prevention efforts if risk factors can be identified. (Pediatr Dent 16:200–5, 1994)

Introduction

There have been significant changes in the epidemiology of coronal tooth caries in the past decade. For children aged 5–17, the mean DMFS was 36% lower in a US nationwide survey in 1986–87[1] than in a similar survey in 1979–80.[2] Despite this decrease, dental caries continues to be a significant health problem for many children. In 1987, acute oral health conditions alone accounted for 3.5 million restricted activity days and nearly a half-million lost school days for children aged 5–17.[3] This should not be surprising because the 1986–87 data reveal that more than half of the children aged 5–17 have dental caries in permanent teeth.

The epidemiology of dental caries in the primary dentition of children has received less attention and investigation than that of permanent teeth. Comparing the 1986–87[1] and 1979–80[2] national surveys, the decline in dmfs for children aged 5–9 was only 26% versus a 36% decline in the DMFS of 12-year-olds. Based on these data it can be concluded that the prevalence of caries in the primary dentition was not only greater than that in the permanent dentition, but its decline from 1979–80 to 1986–87 was less dramatic.

The epidemiology of primary tooth caries needs more exploration. Underscoring this point further is the fact that the actual distribution of primary tooth caries within the US population is unknown. The National Preventive Dentistry Demonstration Project (NPDDP) found that 60% of the dental caries occurred in 20% of the children.[4] This distributional index is quoted often; however it should be noted that these caries distributional data apply to permanent teeth only, not primary teeth. A high-risk group of children with primary tooth caries has not been identified nor characterized; that was a major aim of our study. This information is essential to determine whether preventive methods should be applied universally to all children or targeted to those with elevated risk. In a time of limited oral health care resources, these questions have important health policy implications.

Data on primary tooth caries

The limited information available on primary tooth caries comes from studies in primary dentitions of narrowly defined population groups in developing or industrialized countries.[5] Data from these studies have not been collected by standardized methods, making comparisons very difficult. Moreover, most of these studies have been cross-sectional, so estimates of caries incidence are impossible to infer.

A few studies have attempted to estimate changes in primary tooth caries by comparing caries prevalence in successive cross-sectional studies. A decrease in prevalence of primary tooth caries has been reported from cross-sectional studies of national samples of US children from 1963 to 1987 (Table 1).

Outside the United States, a decrease in primary tooth caries has not been demonstrated clearly. Truin, et al. studied children in The Hague, The Netherlands,

Table 1. Prevalence of dental caries in primary teeth of children in US population studies

Date	*Age*	*deft*	*dfs*
1963–65[7]	6–11	3.0	n.r.
1971–74[7]	6–11	2.7	n.r.
1979–80[2]	5–9	2.6	5.3
1986–87[1]	5–9	1.9	3.9

n.r. = not reported.

- **What are the most relevant publications written by an author?**
- Author-driven query
- Publications are ranked by
  - number of received citations
  - Date
  - popularity (e.g., number of reads)

Discovering cross-topic collaborations among researchers

SCIENTIFIC ARTICLE

Distribution of primary tooth caries in first-grade children from two nonfluoridated US communities

Maureen Quirk Margolis, DDS, MS Ronald J. Hunt, DDS, MS
William F. Vann Jr., DMD, MS, PhD Paul W. Stewart, PhD

Abstract

In a prospective longitudinal study, 1099 first grade children from Aiken, South Carolina, and 1386 children from Portland, Maine, were examined annually for 3 years. Caries prevalence and dmfs incidence were determined. The mean dmfs in Portland children was 2.9. In Aiken, white children had a mean dmfs of 8.4, and black children had a mean dmfs of 10.2. The mean 3-year primary tooth caries increment was 1.5 surfaces in the Portland cohort 3.3 surfaces in the Aiken white cohort and 2.8 surfaces in the Aiken black cohort. These increments were divided evenly between interproximal and fissure surfaces. Twenty percent of the children in Portland had 75% of the caries; in Aiken, 20% of the children had 60% of the caries. This distribution suggests a high-risk group that could be targeted for aggressive caries prevention efforts if risk factors can be identified. (Pediatr Dent 16:200–5, 1994)

Introduction

There have been significant changes in the epidemiology of coronal tooth caries in the past decade. For children aged 5–17, the mean DMFS was 36% lower in a US nationwide survey in 1986–87[1] than in a similar survey in 1979–80.[2] Despite this decrease, dental caries continues to be a significant health problem for many children. In 1987, acute oral health conditions alone accounted for 3.5 million restricted activity days and nearly a half-million lost school days for children aged 5–17.[3] This should not be surprising because the 1986–87 data reveal that more than half of the children aged 5–17 have dental caries in permanent teeth.

The epidemiology of dental caries in the primary dentition of children has received less attention and investigation than that of permanent teeth. Comparing the 1986–87[1] and 1979–80[2] national surveys, the decline in dmfs for children aged 5–9 was only 26% versus a 36% decline in the DMFS of 12-year-olds. Based on these data it can be concluded that the prevalence of caries in the primary dentition was not only greater than that in the permanent dentition, but its decline from 1979–80 to 1986–87 was less dramatic.

The epidemiology of primary tooth caries needs more exploration. Underscoring this point further is the fact that the actual distribution of primary tooth caries within the US population is unknown. The National Preventive Dentistry Demonstration Project (NPDDP) found that 60% of the dental caries occurred in 20% of the children.[4] This distributional index is quoted often; however it should be noted that these caries distributional data apply to permanent teeth only, not primary teeth. A high-risk group of children with primary tooth caries has not been identified nor characterized; that was a major aim of our study. This information is essential to determine whether preventive methods should be applied universally to all children or targeted to those with elevated risk. In a time of limited oral health care resources, these questions have important health policy implications.

Data on primary tooth caries

The limited information available on primary tooth caries comes from studies in primary dentitions of narrowly defined population groups in developing or industrialized countries.[5] Data from these studies have not been collected by standardized methods, making comparisons very difficult. Moreover, most of these studies have been cross-sectional, so estimates of caries incidence are impossible to infer.

A few studies have attempted to estimate changes in primary tooth caries by comparing caries prevalence in successive cross-sectional studies. A decrease in prevalence of primary tooth caries has been reported from cross-sectional studies of national samples of US children from 1963 to 1987 (Table 1).

Outside the United States, a decrease in primary tooth caries has not been demonstrated clearly. Truin, et al. studied children in The Hague, The Netherlands,

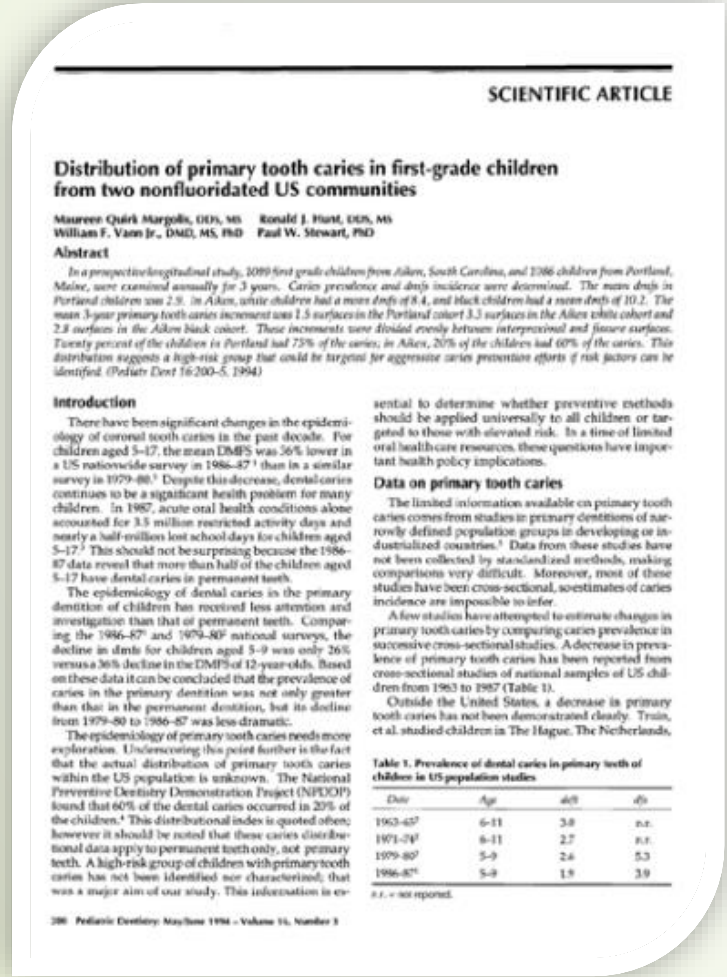
Table 1. Prevalence of dental caries in primary teeth of children in US population studies

Date	*Age*	*deft*	*dfs*
1963–65[7]	6–11	3.0	n.r.
1971–74[7]	6–11	2.7	n.r.
1979–80[2]	5–9	2.6	5.3
1986–87[1]	5–9	1.9	3.9

n.r. = not reported.

- **What are the most relevant publications written by an author on a specific topic?**
- Author- and topic-driven query
- The author's publications covering the topic under analysis are selected and ranked

Discovering cross-topic collaborations among researchers



SCIENTIFIC ARTICLE

Distribution of primary tooth caries in first-grade children from two nonfluoridated US communities

Maureen Quirk Margolis, DDS, MS Ronald J. Hunt, DDS, MS
William F. Vann Jr., DMD, MS, PhD Paul W. Stewart, PhD

Abstract

In a prospective longitudinal study, 1099 first grade children from Aiken, South Carolina, and 1386 children from Portland, Maine, were examined annually for 3 years. Caries prevalence and dmfs incidence were determined. The mean dmfs in Portland children was 2.5. In Aiken, white children had a mean dmfs of 8.4, and black children had a mean dmfs of 10.2. The mean 3-year primary tooth caries increment was 1.5 surfaces in the Portland cohort 3.3 surfaces in the Aiken white cohort and 2.8 surfaces in the Aiken black cohort. These increments were divided evenly between interproximal and fissure surfaces. Twenty percent of the children in Portland had 75% of the caries; in Aiken, 20% of the children had 60% of the caries. This distribution suggests a high-risk group that could be targeted for aggressive caries prevention efforts if risk factors can be identified. (Pediatr Dent 16:200–5, 1994)

Introduction

There have been significant changes in the epidemiology of coronal tooth caries in the past decade. For children aged 5–17, the mean DMFS was 36% lower in a US nationwide survey in 1986–87[1] than in a similar survey in 1979–80.[2] Despite this decrease, dental caries continues to be a significant health problem for many children. In 1987, acute oral health conditions alone accounted for 3.5 million restricted activity days and nearly a half-million lost school days for children aged 5–17.[3] This should not be surprising because the 1986–87 data reveal that more than half of the children aged 5–17 have dental caries in permanent teeth.

The epidemiology of dental caries in the primary dentition of children has received less attention and investigation than that of permanent teeth. Comparing the 1986–87[1] and 1979–80[2] national surveys, the decline in dmfs for children aged 5–9 was only 26% versus a 36% decline in the DMFS of 12-year-olds. Based on these data it can be concluded that the prevalence of caries in the primary dentition was not only greater than that in the permanent dentition, but its decline from 1979–80 to 1986–87 was less dramatic.

The epidemiology of primary tooth caries needs more exploration. Underscoring this point further is the fact that the actual distribution of primary tooth caries within the US population is unknown. The National Preventive Dentistry Demonstration Project (NPDDP) found that 60% of the dental caries occurred in 20% of the children.[4] This distributional index is quoted often; however it should be noted that these caries distributional data apply to permanent teeth only, not primary teeth. A high-risk group of children with primary tooth caries has not been identified nor characterized; that was a major aim of our study. This information is essential to determine whether preventive methods should be applied universally to all children or targeted to those with elevated risk. In a time of limited oral health care resources, these questions have important health policy implications.

Data on primary tooth caries

The limited information available on primary tooth caries comes from studies in primary dentitions of narrowly defined population groups in developing or industrialized countries.[5] Data from these studies have not been collected by standardized methods, making comparisons very difficult. Moreover, most of these studies have been cross-sectional, so estimates of caries incidence are impossible to infer.

A few studies have attempted to estimate changes in primary tooth caries by comparing caries prevalence in successive cross-sectional studies. A decrease in prevalence of primary tooth caries has been reported from cross-sectional studies of national samples of US children from 1963 to 1987 (Table 1).

Outside the United States, a decrease in primary tooth caries has not been demonstrated clearly. Truin, et al. studied children in The Hague, The Netherlands,

Table 1. Prevalence of dental caries in primary teeth of children in US population studies

Date	*Age*	*deft*	*dfs*
1963–65[7]	6–11	3.0	n.r.
1971–74[7]	6–11	2.7	n.r.
1979–80[7]	5–9	2.6	5.3
1986–87[1]	5–9	1.9	3.9

n.r. = not reported.

200 Pediatric Dentistry: May/June 1994 – Volume 16, Number 3

- **What are the most fruitful collaborations among multiple authors?**
- No deterministic solution
- Hard to solve using simple queries
  - For each topic?
  - For each combination of authors?
  - How to combine and rank the results?

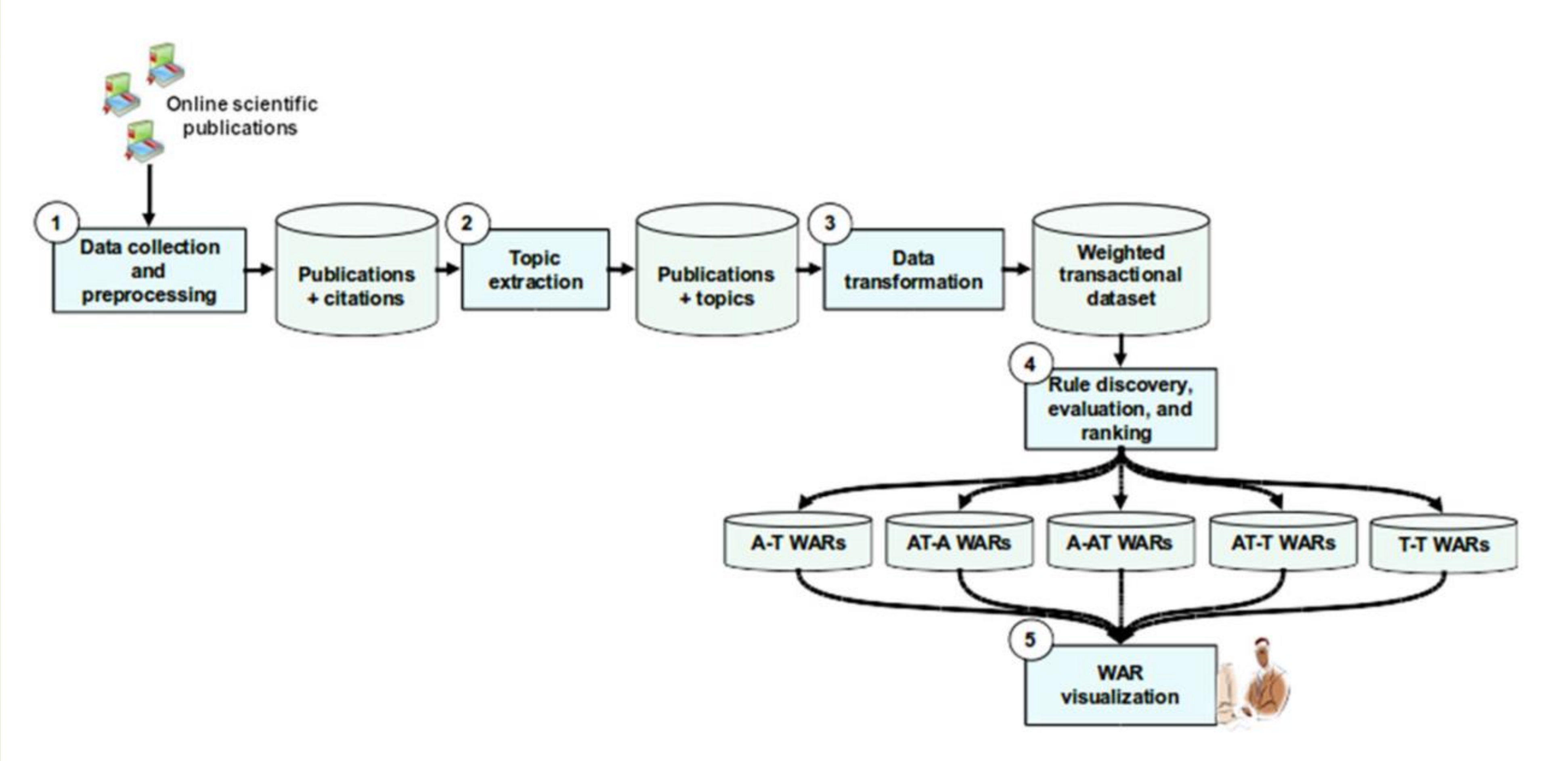
Discovering cross-topic collaborations among researchers

- Identify fruitful collaborations among researchers
  - By analyzing co-authored scientific publications and their popularity/relevance in terms of number of citations

- Expected result (automatically inferred from publications data):
  - the discovery of research collaborations among multiple authors on single or multiple topics.

- The main novelty is the fact that we are able to extract correlation between set of authors and topics
  - Previous approaches usually focused only on the correlations between single author and a topic

Cross-topic Scientific Collaboration Analyzer



Data collection

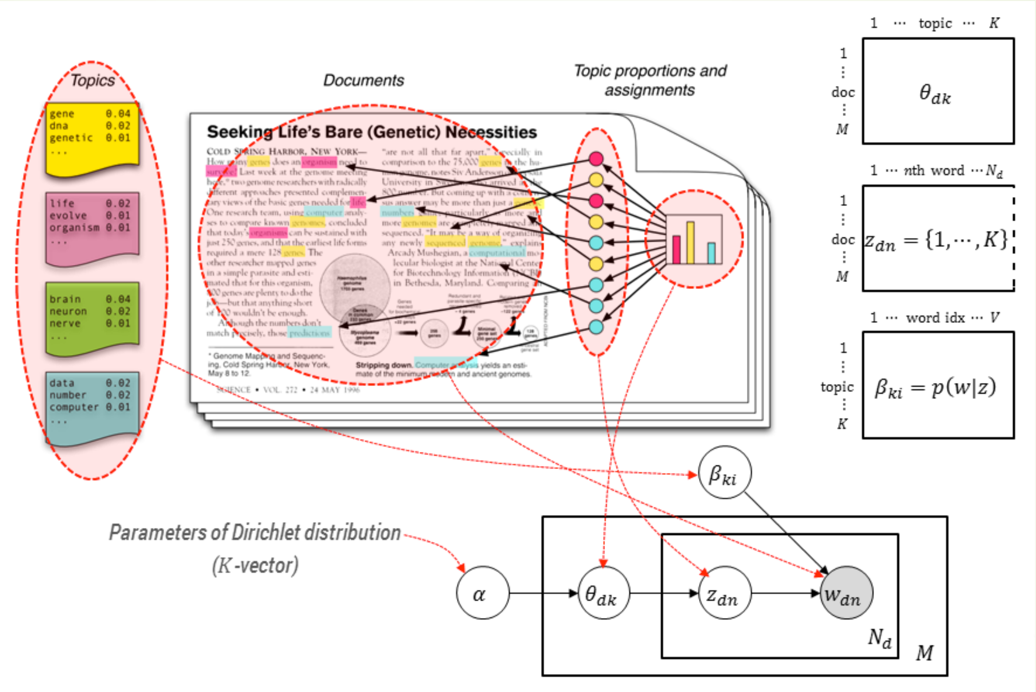
- Publication data are acquired from digital libraries and online databases
  - e.g., PubMed (NCBI 2017), OMIM (Hamosh et al. 2000)
    - by exploiting the exposed Application Programming Interfaces (APIs) and then stored in a unique repository.

For each publication we acquire the following data:

- 1. the Digital Object Identifier (DOI) of the publication,
- 2. the list of authors,
- 3. the current number of citations received,
- 4. the text of the publication, and
- 5. any relevant (domain-specific) metadata associated with the publication

- The current number of citations is considered as one of the main indicators of influence/popularity of a scientific publication in the research community

Topic extraction



- We proposed two complementary strategies to assign topics to each publication:
  1. if topic metadata are given, CSCA exploits **metadata** content as descriptors of the covered topic.
  2. Otherwise, from the textual content of the publication
     - ✓ Then we exploited the **Author-Topic Model (ATM)** (Rosen-Zvi et al. 2012)
       - ❖ To extract topic
       - ❖ top-K topics for each document will return

Topic extraction: ATM topics

Topic ID	Top-10 most related terms
T0	rat, neuron, muscl, effect, dai, studi, calcium, group, activ, induc
T1	gene, mutat, express, sequenc, protein, develop, analysi, dna, cell, genet, genom
T2	respons, drug, increas, potenti, channel, membran, effect, studi, function, reduc
T3	cancer, associ, studi, breast, increas, case, model, genotyp, risk, smoke
T4	health, data, base, method, studi, model, system, develop, predict, approach
T5	brain, imag, memori, tissu, inject, studi, model, control, test, network
T6	infect, hiv, viru, associ, immun, vaccin, diseas, antigen, reactiv, hepat
T7	cell, express, activ, induc, tumor, human, regul, protein, mice, receptor
T8	protein, activ, cell, bind, fig, membran, acid, level, α, dna
T9	patient, studi, group, ag, risk, conclus, year, method, treatment, associ

Data transformation

- Weighted transactional dataset
  - Set of weighted transactions
  - Each transaction represents a different publication and consists of a set of items
    - Items are either authors or topics
  - Transactions are weighted by a relevance weight (e.g., the number of received citations)

Pub. id	**#cit.**	**Authors**	**Topics**
1	10	$(Author : Brown,\ J.), (Author : Smith,\ L.)$	$(Topic : A), (Topic : X)$
2	5	$(Author : Brown,\ J.), (Author : Smith,\ L.)$	$(Topic : D), (Topic : X)$
3	10	$(Author : Brown,\ J.), (Author : Smith,\ L.)$	$(Topic : C), (Topic : Z)$
4	1	$(Author : Smith,\ L.)$	$(Topic : X), (Topic : Z)$
5	10	$(Author : Brown,\ J.), (Author : Smith,\ L.)$	$(Topic : C)\ (Topic : X)$
6	12	$(Author : Smith,\ L.)$	$(Topic : Z)$

The pattern-based solution

- Pattern mining
  - Apply a weighted itemset mining algorithm to extract frequent patterns from the weighted dataset
    - The weight of each extracted itemset is given by the sum of the relevance weights of the papers associated with that itemset
    - **weighted support** index represents the weighted frequency of occurrence of the rule in the source dataset
    - **weighted confidence** index represents the rule strength.
  - Focus on the frequent patterns representing correlations between authors and topics
    - Authors – Topic Patterns (ATP)
      - E.g., {(Author: Smith L.), (Author: Johnson A.),(Topic, Z)}

The pattern-based solution (example)

- WAR {(Author:Brown; J:),(Author:Smith; L:)} ⟶ (Topic : X)}
  - indicates an implication between a couple of authors and a specific topic.
  - weighted support equal to 25
  - weighted confidence equal to $\frac{25}{35}$

Pub. id	#cit.	Authors	Topics
1	10	$(Author : Brown,\ J.), (Author : Smith,\ L.)$	$(Topic : A), (Topic : X)$
2	5	$(Author : Brown,\ J.), (Author : Smith,\ L.)$	$(Topic : D), (Topic : X)$
3	10	$(Author : Brown,\ J.), (Author : Smith,\ L.)$	$(Topic : C), (Topic : Z)$
4	1	$(Author : Smith,\ L.)$	$(Topic : X), (Topic : Z)$
5	10	$(Author : Brown,\ J.), (Author : Smith,\ L.)$	$(Topic : C)\ (Topic : X)$
6	12	$(Author : Smith,\ L.)$	$(Topic : Z)$

Weighted association rules categories

1. Authors-Topic (A-T) Rules
2. AuthorsTopic-Author (AT-A) Rules
3. Authors-AuthorTopic (A-AT) Rules
4. AuthorsTopics-Topic (AT-T) Rules
5. Topics-Topic (T-T) Rules

1) Authors-Topic Rules

- On what topics is the collaboration focused on?
- Is the collaboration focused on a specific topic or spread over multiple topics?
- For example: **{(Author:Brown; J:),(Author:Smith; L:)}** → **(Topic : X)} is an A-T**
  - It indicates that authors J. Brown and L. Smith have co-authored publications related to topic X.
- The weighted support indicates the sum of the citation counts of all the co-authored publications on the given topic.
- A-T WARs with high weighted confidence indicate the topics on which the collaboration is mainly focused on.
- For example, if the wconf of a A-T WAR is close to 100% (all the citations are associated with a particular topic)
  - it means that the collaborations was productive only on the corresponding topic.

2) AuthorsTopic-Author Rules

- Working on a given set of topics, has the group (occasionally) collaborated with external authors?
- This rule indicates the significance of the collaboration between the group under analysis and the external author.
- **For example: {(Author:Brown; J:),(Author:Smith; L:), (Topic:X)} → (Author:Black; J:)**
  - indicates that in the collaboration between authors J. Brown and L. Smith on topic X they have collaborated with author J. Black.
- The weighted support indicates the significance of the collaboration between the group under analysis and the external author.
- The weighted confidence indicates the impact of this collaboration on the productivity of the group of authors associated with the given topic.
  - low wconf value indicate occasional (yet potentially fruitful) collaborations
  - high wconf values indicate more systematic collaborations between the group and external authors
- For example, if the wconf is 50%
  - it means that half of the citations received by the combination of authors on the considered topic was achieved by works co-authored by the considered author.

3) Authors-AuthorTopic Rules

- Has the group collaborated with external authors? On which topics?
- indicates the significance of the collaboration between the group of authors and the consider pair author-topic
- **For example: {(Author:Brown; J:),(Author:Smith; L:)} → {(Author:Black; J:), (Topic : X)}**
  - indicates that in the research works made in the collaboration between authors J. Brown and L. Smith the authors have frequently collaborated with author J. Black on topic X.
- The weighted confidence indicates the impact of this topic-specific collaboration on the overall productivity of the group of authors in the antecedent of the rule (independently of the topic).
  - For example,
    - if the wconf is 50% it means that half of the citations received by the combination of authors (independently of the topic) was achieved by works co-authored by the external author on the indicated topic.
    - Low wconf values may be due either to the low productivity of the collaboration between the group and the external authors or to the low popularity of the topic

4) AuthorsTopics-Topic Rules

- Given a group of researchers who have frequently collaborated on a set of topics, which other topic is likely to be covered by their coauthored publications?
- describe cross-collaborations between authors.
  - Since in a collaboration each member could provide its expertise on a particular topic, it is interesting to investigate on which topics an existing author-topic collaboration could be specialized.
- **For example, {(Author:Brown; J:), (Author:Smith; L:), (Topic : X)} → {(Topic : Z)}**
  - indicates that an authors' collaboration on topic X is frequently associated with an additional topic (Z).
- If the wconf of the AT-T WAR is very high (close to 100%)
  - most of the co-authored publications related to topic X cover topic Z as well.

5) Topics-Topic Rules

- To which topic is a particular set of topics most correlated with?
- Since authors' collaborations are often cross-topic, analyzing the underlying correlation between multiple topics is particularly interesting.
- **For example, {(Topic : A), (Topic : X)} → {(Topic : Z)}**
- Sorting T-T WARs by decreasing confidence
    - allows us to identify the sets of most correlated sets of topics

Visualization

Rules

[A-Authors:Daniel_S.E., A-Authors:Kilford_L., A-Authors:Lees_A.J., D-Disease:Topic3, D-Disease:Topic5, D-Disease:Topic6, D-Disease:Topic8, D-Disease:Topic7]->[A-Authors:Hughes_A.J.],Support=1087,Confidence=1.0

[A-Authors:Hughes_A.J., A-Authors:Daniel_S.E., A-Authors:Kilford_L., D-Disease:Topic3, D-Disease:Topic5, D-Disease:Topic6, D-Disease:Topic8, D-Disease:Topic7]->[A-Authors:Lees_A.J.],Support=1087,Confidence=1.0

[A-Authors:Hughes_A.J., A-Authors:Daniel_S.E., A-Authors:Lees_A.J., D-Disease:Topic3, D-Disease:Topic5, D-Disease:Topic6, D-Disease:Topic8, D-Disease:Topic7]->[A-Authors:Kilford_L.],Support=1087,Confidence=1.0

[A-Authors:Hughes_A.J., A-Authors:Kilford_L., A-Authors:Lees_A.J., D-Disease:Topic3, D-Disease:Topic5, D-Disease:Topic6, D-Disease:Topic8, D-Disease:Topic7]->[A-Authors:Daniel_S.E.],Support=1087,Confidence=1.0

[A-Authors:Buxton_J., D-Disease:Topic6, D-Disease:Topic8]->[A-Authors:Johnson_K.],Support=456,Confidence=1.0

[A-Authors:Johnson_K., D-Disease:Topic6, D-Disease:Topic8]->[A-Authors:Buxton_J.],Support=456,Confidence=1.0

[A-Authors:Harley_H.G., A-Authors:Shaw_D.J., D-Disease:Topic6]->[A-Authors:Harper_P.S.],Support=413,Confidence=1.0

[A-Authors:Harper_P.S., A-Authors:Harley_H.G., D-Disease:Topic6]->[A-Authors:Shaw_D.J.],Support=413,Confidence=1.0

[A-Authors:Harper_P.S., A-Authors:Shaw_D.J., D-Disease:Topic6]->[A-Authors:Harley_H.G.],Support=413,Confidence=1.0

[A-Authors:Harley_H.G., A-Authors:Shaw_D.J., D-Disease:Topic9, D-Disease:Topic6, D-Disease:Topic0]->[A-Authors:Harper_P.S.],Support=408,Confidence=1.0

Real case scenario

Top 5 Authors-Topic rules (A-T WARs) in terms of wsup

A-T rule	wsup	wconf (%)
(*Author*:Siddique, T.), (*Author*:Deng, H.-X.) → (*Topic*:AMYOTROPHIC LATERAL SCLEROSIS 1)	1828	100
(*Author*:Hentati, A.), (*Author*:Siddique, T.), (*Author*:Deng, H.-X.) → (*Topic*:AMYOTROPHIC LATERAL SCLEROSIS 1)	1800	100
(*Author*:Rioux, J.D.), (*Author*:Silverberg, M.S.) → (*Topic*:INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1)	1470	100
(*Author*:Silverberg, M.S.), (*Author*:Barmada, M.M.) → (*Topic*:INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1)	1388	100

- authors Siddique T. and Deng H. X. wrote a set of papers on the Amyotrophic lateral sclerosis
- their co-authored publications have been cited 1828 times.
- this WAR is the most frequent one among all the mined A-T WARs ranging over the topic
  - we can deduce that Siddique T. and Deng H. X. are among the most influential/authoritative group of researchers about Amyotrophic lateral sclerosis.

Future work

- What are the most appropriate objective and subjective measures to evaluate the interestingness of a rule? How can we effectively drive the user exploration of the mined rules?
  - we envision the integration in the proposed methodology of more advanced rule quality indices
  - we aim at collecting the user relevance feedbacks on the mined rules by enriching the Web-based interface
    - These feedback scores can be exploited to enhance the quality of the generated model or to refine the process of rule generation based on users' preferences.
- Application to Reviewer Assignment Problem
  - in the peer reviewing process academic papers are assigned to anonymous reviewers with complementary expertise to assess the innovative contribution of their submitted work

Thank you for your attention

Questions?