ScuDo
Scuola di Dottorato ∽ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (31^{st} cycle)

Data Mining and Indexing Big Multimedia Data

By

Mohammad Reza Kavoosifar

******

Supervisor:
Prof. Elena Baralis

Doctoral Examination Committee:
Prof. Giuseppe Psaila, Università di Bergamo
Prof. Elisa Quintarelli, Università di Verona

Politecnico di Torino
2019

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Mohammad Reza Kavoosifar
2019

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Acknowledgements

I would like to acknowledge the guidance of my supervisor, Prof. Elena Baralis, and her support during the years of the PhD. A big part of this thesis would not be accomplished without the precious help and the experience of Prof. Paolo Garza, to whom I am deeply thankful. My sincere thanks also go to Prof. Benoit Huet, my research would have been impossible regardless the aid and support of him.

I am grateful to all of those with whom I have had the pleasure to work during this and other related projects.

Nobody has been more important to me in the pursuit of this research than the members of my family. I would like to thank my parents and my brother, whose love and guidance are with me in whatever I pursue. They are the ultimate role models.

Abstract

Increasing amounts of multimedia content are being produced and stored on a daily basis. In order to make this data useful, computer applications are required that facilitate search, browsing, and navigation through these large data collections.

The first part of this thesis describes our approach to carry out multimedia search and indexing by connecting the textual information and visual content. The experiments were carried out considering the TRECVID Video Hyperlinking task.

Different combinations of monomodal queries are experimentally evaluated, and the impact of both parameters and single features are discussed to identify their contributions. The Automatic Feature Selection (AFS) algorithm gain the best-performing approach at the TRECVID 2017 video hyperlinking challenge. The proposed algorithm includes three different monomodal queries based on enriched feature sets.

The second part of this thesis is related to textual information analysis for discovering of research collaborations among multiple authors on single or multiple topics. Identifying the most relevant scientific publications on a given topic is a well-known research problem. The Author-Topic Model (ATM) is a generative model that represents the relationships between research topics and publication authors. It allows us to identify the most important authors on a particular topic. Specifically, we exploited an exploratory data mining technique, i.e., Weighted Association Rule (WAR) mining, to analyze publication data and to discover correlations between ATM topics and combinations of authors.

The applicability of the proposed approach was validated on real data acquired from the Online Mendelian Inheritance in Man catalog of genetic disorders and from the PubMed digital library. The results confirm the effectiveness of the proposed strategy.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

During my PhD, I used data mining and machine learning techniques to solve two different problems. The first one is more related to Multimedia data. Although, the content and the original raw data is multimedia documents but I focused my research on the textual information associated with these multimedia documents.

With the explosive growth and the widespread accessibility of multimedia content, video content is becoming one of the most valuable sources to assess information and knowledge. While watching a video, it is common that users are interested in finding further information on some aspects of the topic of interest contained within a video segment. Therefore, it is crucial to develop effective video search and hyperlinking techniques to help users explore, navigate and search video contents in audiovisual archives.

Unfortunately, relevance of similar content in terms of textual and visual concepts does not offer diversity in the set of results.This lack of diversity is considered as destructive in many exploration scenarios. For this reason, it would be a desirable idea by providing relevant links for covering a number of possible extensions with respect to the anchor's content. Specifically, a set of diverse results is required in order to improve the chance of any user to find at least one interesting link to follow. This objective of providing diverse results is an important goal in the hyperlinking analysis.

Video hyperlinking consists in linking a video anchor or segment to other video segments in a video collection, based on similarity or relatedness. Accordingly, video hyperlinking enables users to navigate between video segments in a source content.

For this analysis, we applied some database techniques specifically indexing techniques on the textual part of data in order to identify relevant video segments. In this part, we described the framework used by the Eurecom-Polito team [1, 2] to address the Hyperlinking task inside a video collection at TRECVID 2017 [3]. We have proposed a system that exploits different combinations of monomodal queries. Each query is based on textual features, enriched with concepts and entities aimed at maximizing the relevance of the selected video segments. The exploited features are: (i) automatic speech transcripts [4, 5], (ii) visual concepts, (iii) entities extracted by Named-Entity Recognition techniques, and (iv) a concept mapping technique, which is based on WordNet [6].

My second topic is more focused on textual documents related to research papers. Specifically, I have exploited a set of pattern mining techniques that allowed me to address specific problems such as author-topic identification.

Most scientific publications like conference proceedings, scientific journal, and books are accessible through digital libraries and online databases. For example, in genetics and genomics PubMed [7] and OMIM [8] are among the most popular publication repositories. Researchers generally perform manual topic- or author-driven queries on publication data to retrieve the content of interest. However, this activity can be extremely time consuming and susceptible to errors, as a result of the amount of publications to explore is also large.

In this thesis, an analysis has been performed on the problem of discovering cross-topic collaborations among multiple authors by means of an exploratory data mining technique, i.e., weighted association rule mining [9].

The effectiveness of the proposed approach has been experimentally evaluated on data acquired from two independent libraries, i.e., OMIM [8] and PubMed [7], which collect genomic and genetic studies.

The rest of this thesis is organized as follows.

Chapter 2 describes the hyperlinking problem and the proposed algorithms. In this chapter, Section 2.1 provides characterization of video data. Section 2.2 presents related works. Section 2.3 introduces the proposed system and its main phases. Section 2.4 provides details into the query formulation phase. Section 2.5 describes the different combinations of features exploited to retrieve relevant video segments. Section 2.6 presents and discusses the experimental results obtained by the proposed combinations on the TRECVID 2017 dataset. Section 2.7 provides details regarding the evaluation of concepts detected by the visual concept detector.

Chapter 3 describes the problem related to discovering collaborations among authors. In this chapter, Section 3.1 compares the proposed approach with existing studies. Section 3.2 thoroughly describes the proposed methodology, while Section 3.3 experimentally evaluates its effectiveness on real data.

Finally, Chapter 4 discusses conclusions of the thesis.

Chapter 2

Video hyperlinking

Value from the rapidly growing archives of produced digital multimedia content will only be realized with the development of technologies that allow users to explore them through search and retrieval of potentially interesting content.

The Video Hyperlinking aims at linking anchors related to a temporal segment of a video. In this task, one of the main challenges is the uncertainty regarding what criteria are to be followed to generate these links. 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.

Figure 2.1 is an example picture giving an impression of video hyperlinking in a video segment on tourism in London: an item on a Fish & Chips restaurant could be linked to a cooking program describing a recipe for Fish & Chips, an item on the London Parliament could be linked to segments about England's Queen.

Relevance of a link target can be based upon topical information, the events or activities depicted, the people present in the videos, etc. However, finding similar target video segments given an anchor video segments is not the aim in video hyperlinking.

2.1 Video data characterization

Video content structuring

Based on Figure 2.2, a video can be structured in a hierarchical form.

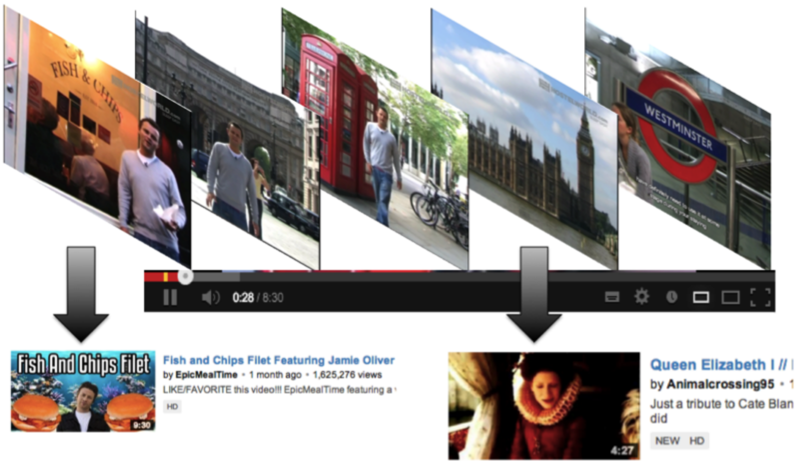

FISH & CHIPS
WESTMINSTER
0:28 / 8:30
Fish And Chips Filet
9:30
Fish and Chips Filet Featuring Jamie Oliver
by EpicMealTime • 1 month ago • 1,625,276 views
LIKE/FAVORITE this video!!! EpicMealTime featuring a
HD
4:27
Queen Elizabeth I //
by Animalcrossing95 •
Just a tribute to Cate Blan
did
NEW HD

Fig. 2.1 An example to video Hyperlinking [10]

- a **scene** is defined as a collection of semantically related and temporally adjacent shots
- a **shot** is an uninterrupted clip recorded by a single camera. It is a physical entity which often forms the building block of video content
- a **keyframe** is the frame which best represents the content of a shot

Anchor

An anchor is a video segment that a user is currently watching, which is defined by a start and an end time within a video.

Visual concepts

Visual concepts are the concepts which are being detected in a keyframe by exploiting an image processing tool. In this research, We used only the text (name) of these concept.

2.2 Related work

The automatic generation of hyperlinks within video collections has recently become a major subject, specifically in some evaluation benchmarks such as MediaEval and TRECVID [12, 13]. The key idea is to create hyperlinks between video segments within a collection, enriching a set of anchors that represent interesting entry points

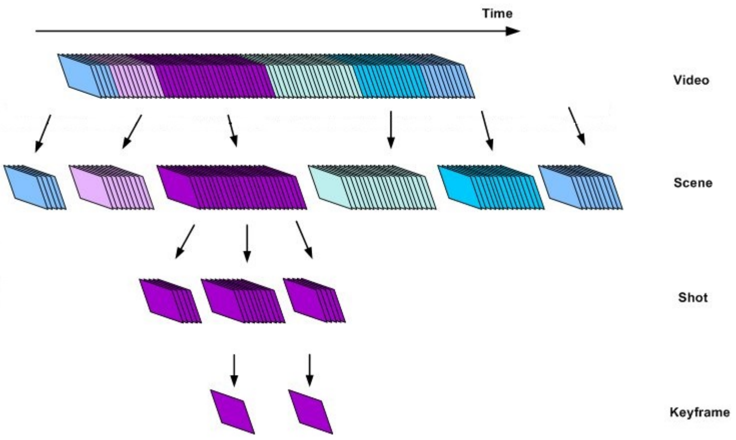


Fig. 2.2 Video content structuring [11]

in the collection itself. Links can be seen as recommendations for potential viewers, whose intent is not known at the time of linking. The goal of the links is thus to help viewers gain insights on a potentially massive collection of videos so as to find information of interest, following a search and browse paradigm. To this aim, several techniques have been proposed.

Besides the unimodal approaches, such as [14], which relies on textual features only, multi-modal techniques taking into account different feature sets have emerged. In [15], Soleymani et al. have proposed a multi-modal system designed to analyze users' behavior and interaction with browsed visual content for different image search intents, whereas the approach proposed in our paper exploits combinations of many different features, both textual and visual.

Additional paradigms propose models predominantly based on one specific modality (e.g., image search) and try to improve them using information from other modalities (e.g., captions) [16, 17]. Similarly, [18, 19] propose a text-to-video mapping. On the other hand, [20] described a system for content-based video retrieval from large surveillance video archives, using behavior, actions and appearance of objects. Recent high-performing approaches in video browsing revolve around retrieval of simplified sketches (e.g., by using simple color signatures [21]) and displaying the collection in a more informative way (e.g., using a graph-based keyframe arrangement for browsing [22]). A more in-depth sketch analysis where deep semantic classifiers are employed for sketch auto-completion has been demonstrated also in an earlier work [23]. A vertical application of hyperlink techniques is presented in [24], where an effective signature-based approach has been proposed to link endoscopic images with video segments.

A new indexing and retrieval system is presented in [25]. It detects multiple object events or crowd events (e.g., group walking, group splitting, etc.). However, the generic video hyperlinking use case requires not only the detection of group items, but also single items or objects which are appearing inside the videos.

Some other approaches are also developed in Multimodal Video Retrieval. The IMOTION system [26] represents a multimodal content-based video search and browsing application offering a rich set of query modes based on a feature-fusion approach. The VERGE interactive search engine [27] is capable of browsing and searching into video content by providing integrated content-based analysis and retrieval modules, such as video shot segmentation, concept detection, clustering, and visual-similarity and object-based search. In terms of using features, the approach proposed in the current paper exploits a different set of features, for instance by including also video metadata, and by avoiding the need to perform video processing tasks since it relies on textual provided features.

Leveraging different information sources is a task investigated by [28]. They include video and text for efficient video browsing, however, the search and hyperlinking task [12] is to seek for meaningful videos with respect to a text query. Advances have been reported in the area of cross-modal systems by IRISA team [29] and VIREO teams [30]. Cross-modal systems are based on two (or more) modalities that are known to share a common set of categories.

The IRISA group exploited an enriched version of their 2016 algorithm, a crossmodal *Bidirectional Deep Neural Networks (BiDNN)* Joint Learning [31], which ranked first in TRECVID 2016. In their 2016 algorithm [32], training is performed cross-modally and in both directions: one modality is presented as an input and the other as the expected output, and vice-versa at the same time (i.e., the second one is presented as input and the first one as expected output). This is equivalent to using two separate deep neural networks and tying them (sharing specific weight variables) to make them symmetrical. Finally, for the phase of video hyperlinking, segments are compared. For each video segment, the two modalities are considered: embedded automatic transcripts with embedded CNN (a very deep Convolutional Neural Network [33]) representation) and a multimodal embedding is created with a bidirectional deep neural network. Then, the two multi-modal embeddings are compared with a cosine distance to obtain a similarity measure. However, for TRECVID 2017, the IRISA group, contrary to their 2016 algorithm, decided to put more emphasis on the choice of visual descriptors. Additionally, the use of metadata was explored in one of the runs.

The VIREO group introduced a deep model called *Semantic Representation Network (SRN)* which evaluates the relatedness between visual and text data. The structure of SRN contains different layers. At first, it consists of two networks, which share weights with each other, for inputs of anchors and targets. Then it encodes both target and anchor into the same feature space. After that, the holographic layer would evaluate the relatedness between anchor and target by exploiting circular correlation [34], which measures vector correlation in the frequency domain using FFT (Fast Fourier Transform). Finally, the softmax layers output the probabilities of similarity and dissimilarity between anchors and targets. For the phase of video hyperlinking and for their 4 submitted runs at TRECVID 2017, they considered 2 algorithms. For Run-1 (Visual baseline), they exploited SRN and cosine similarity. Then for Run-3 (Multimodal baseline), they combined visual Run-1 and the text features extracted from ASR (Automatic Speech Recognition). For the other 2 runs (Run-2 and Run-4), they formulated the problem as an optimization algorithm (considering k-nearest neighbors) and adopted LID-first algorithm [35] for re-ranking of baseline results. The goal of this algorithm is to promote the ranks of targets with "lower data risk", specifically, in lower local dimensions, being hubs of data, and sufficiently diverse from neighboring regions.

Even if such proposals are all very promising, our approach gained higher MAiSP (Mean Average interpolated Segment Precision) [36] in the TRECVID 2017 workshop, thanks to the proposed combinations of multi-modal features.

Finally, in [37] additional studies on cross-modal systems are presented, however they work well only in terms of text-to-image retrieval, while our approach considers both image-to-text and text-to-image aspects.

2.3 Problem statement and system overview

The video hyperlinking (LNK) task at TRECVid aims to foster progress in tools for effectively accessing video content. The task begins with an anchor video segment. The goal is to produce a ranked list of relevant segments to the anchor.

For the Hyperlinking task, we developed a system based on both textual and visual features. We exploited all the data and metadata provided by the task organizers, except visual concepts. Specifically, we decided to use the visual concepts extracted by using the Caffe framework with the BVLC GoogLeNet model [38]. We also considered some other extra features. Specifically, to identify the more relevant terms and concepts in each query we used the Stanford Named Entity Recognizer (NER) [39] software to find entities and a Concept mapping technique based on WordNet [6].

The proposed system exploits (i) automatic speech recognition transcripts (LIMSI) [4, 5], (ii) visual concepts, based on the Caffe framework, (iii) meta-data of the videos (specifically, title, description and tags have been considered), and (iv) query reformulation (based on Named-entity recognition and Concept mapping).

Overall, the proposed system is based on the following features:

- Automatic speech recognition transcripts (LIMSI) [4, 5].
- Visual concepts, provided by the ImageNet GoogleNet model.
- Metadata of the videos (specifically, title, description, and tags).
- Results of named-entities recognition and concept mapping.

The system exploits a three-step approach, with each step associated to a computation stage, as presented in Figure 2.3:

1. Data segmentation (Section 2.3.2).
2. Indexing (Section 2.3.3).
3. Query formulation and retrieval (Section 2.3.4).

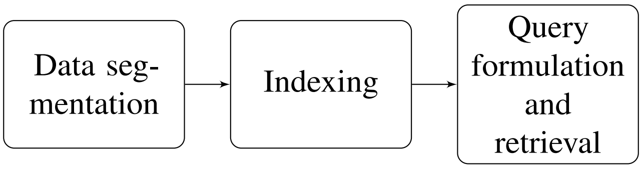


Fig. 2.3 System stages

2.3.1 Data features

For this research, we exploit the video dataset which used for the TRECVID 2017 competition and has been provided by blip.tv (Blip10000 dataset) [40]. It consists of 14,838 videos, for a total of 3,288 hours. The mean length of videos is around 13 minutes. Videos are characterized by metadata (we considered title, short program descriptions, and tags (Figure 2.4)), Automatic Speech Recognition (ASR) transcripts (LIUM and LIMSI), visual concepts, shots, and keyframes.

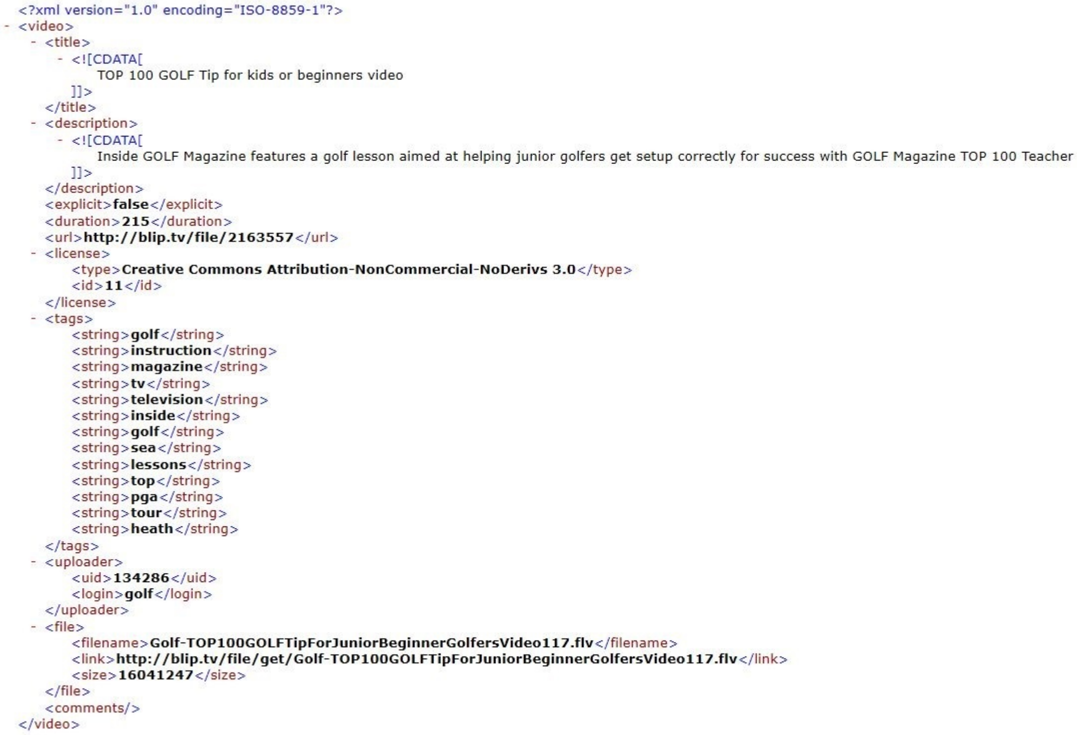
Figure 2.5 shows a sample of LIMSI 2016 transcript, while Figure 2.6 shows a sample of LIUM 2012 transcript for the same video. In both transcript files, the words detected during a speech is reported by start and end time as well as a corresponding score.

Each Visual concept file is formatted in CSV (Figure 2.8) and contains a list of concepts ids that is being detected, along with the corresponding score. These Ids are referenced to the synset words for the concept. Figure 2.9 shows a sample of synset words file of visual concepts.

Figure 2.10 represents a sample keyframe of the video related to the previous data samples.

The videos present a variety of topics from computer science tutorials and sightseeing guides to homemade song covers. They are provided in many languages but a vast majority of them are in English, while the anchor video fragments were exclusively in English.

The training set provided by TRECVID contains 90 query anchors and their corresponding set of ground-truth related segments. The test set consists of 25 different query anchors. Figure 2.7 shows a sample of the training set.



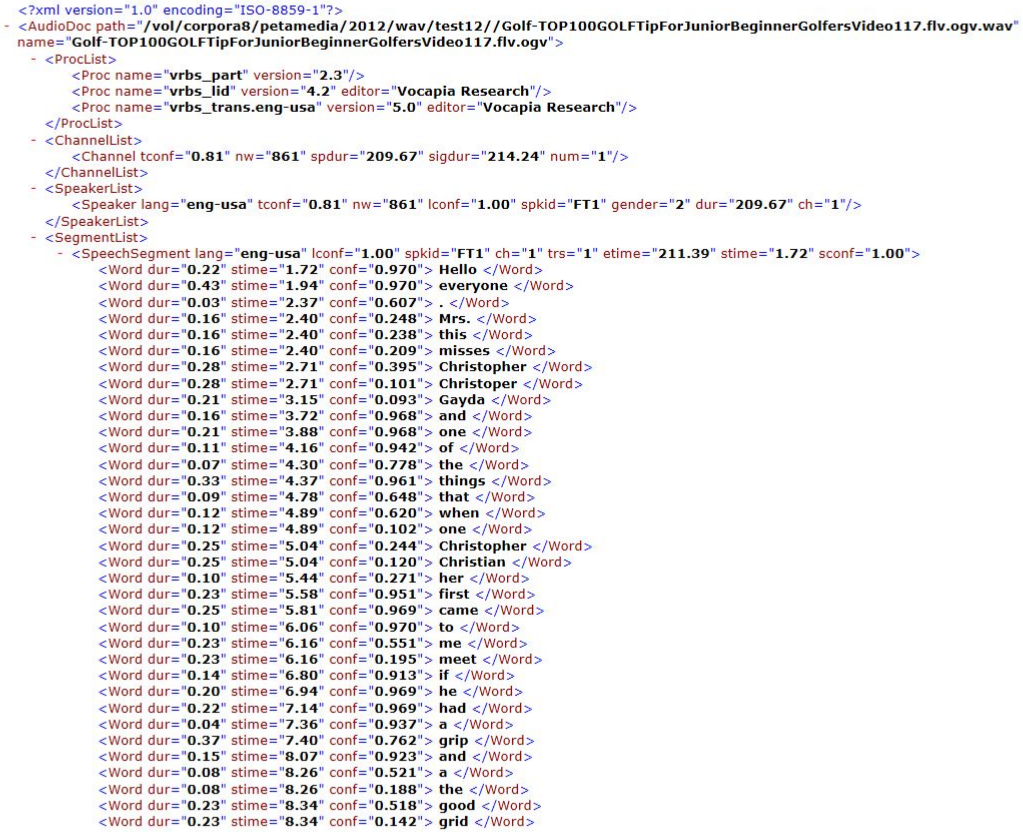
```
  <?xml version="1.0" encoding="ISO-8859-1"?>
- <video>
  - <title>
    - <![CDATA[
        TOP 100 GOLF Tip for kids or beginners video
      ]]>
    </title>
  - <description>
    - <![CDATA[
        Inside GOLF Magazine features a golf lesson aimed at helping junior golfers get setup correctly for success with GOLF Magazine TOP 100 Teacher
      ]]>
    </description>
    <explicit>false</explicit>
    <duration>215</duration>
    <url>http://blip.tv/file/2163557</url>
  - <license>
      <type>Creative Commons Attribution-NonCommercial-NoDerivs 3.0</type>
      <id>11</id>
    </license>
  - <tags>
      <string>golf</string>
      <string>instruction</string>
      <string>magazine</string>
      <string>tv</string>
      <string>television</string>
      <string>inside</string>
      <string>golf</string>
      <string>sea</string>
      <string>lessons</string>
      <string>top</string>
      <string>pga</string>
      <string>tour</string>
      <string>heath</string>
    </tags>
  - <uploader>
      <uid>134286</uid>
      <login>golf</login>
    </uploader>
  - <file>
      <filename>Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv</filename>
      <link>http://blip.tv/file/get/Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv</link>
      <size>16041247</size>
    </file>
    <comments/>
  </video>
```

Fig. 2.4 A sample of Metadata

2.3.2 Data segmentation

The first step that is applied on the video collection consists in splitting the videos in segments. We used a Fixed-segmentation, for which we considered 120 sec fixed segments. Previous year experiments [41] showed that Shot-segmentation is not a good choice to investigate as the videos are a collection of semi-professional user-generated videos where they are not edited and for most of them, people filmed themselves. For this reason, we investigate on Fixed-length segmentation. We chose 120 seconds fixed-segmentation because they seem to provide better coverage and more choice than the lower length segmentation. Also the 120 seconds is the upper bound for an anchor in the Hyperlinking task (the minimum length is 10 seconds).

All the textual data associated with the segments have been preprocessed to remove irrelevant words. Specifically, we used a punctuation removal tool and we also removed stop-words. Stopword elimination filters out the words having least semantic content, because their presence would bias the quality of the next phase. Furthermore, we narrowed down the word list of each segment to its core



```
  <?xml version="1.0" encoding="ISO-8859-1"?>
- <AudioDoc path="/vol/corpora8/petamedia/2012/wav/test12//Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv.wav"
  name="Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv">
  - <ProcList>
        <Proc name="vrbs_part" version="2.3"/>
        <Proc name="vrbs_lid" version="4.2" editor="Vocapia Research"/>
        <Proc name="vrbs_trans.eng-usa" version="5.0" editor="Vocapia Research"/>
    </ProcList>
  - <ChannelList>
        <Channel tconf="0.81" nw="861" spdur="209.67" sigdur="214.24" num="1"/>
    </ChannelList>
  - <SpeakerList>
        <Speaker lang="eng-usa" tconf="0.81" nw="861" lconf="1.00" spkid="FT1" gender="2" dur="209.67" ch="1"/>
    </SpeakerList>
  - <SegmentList>
      - <SpeechSegment lang="eng-usa" lconf="1.00" spkid="FT1" ch="1" trs="1" etime="211.39" stime="1.72" sconf="1.00">
            <Word dur="0.22" stime="1.72" conf="0.970"> Hello </Word>
            <Word dur="0.43" stime="1.94" conf="0.970"> everyone </Word>
            <Word dur="0.03" stime="2.37" conf="0.607"> . </Word>
            <Word dur="0.16" stime="2.40" conf="0.248"> Mrs. </Word>
            <Word dur="0.16" stime="2.40" conf="0.238"> this </Word>
            <Word dur="0.16" stime="2.40" conf="0.209"> misses </Word>
            <Word dur="0.28" stime="2.71" conf="0.395"> Christopher </Word>
            <Word dur="0.28" stime="2.71" conf="0.101"> Christoper </Word>
            <Word dur="0.21" stime="3.15" conf="0.093"> Gayda </Word>
            <Word dur="0.16" stime="3.72" conf="0.968"> and </Word>
            <Word dur="0.21" stime="3.88" conf="0.968"> one </Word>
            <Word dur="0.11" stime="4.16" conf="0.942"> of </Word>
            <Word dur="0.07" stime="4.30" conf="0.778"> the </Word>
            <Word dur="0.33" stime="4.37" conf="0.961"> things </Word>
            <Word dur="0.09" stime="4.78" conf="0.648"> that </Word>
            <Word dur="0.12" stime="4.89" conf="0.620"> when </Word>
            <Word dur="0.12" stime="4.89" conf="0.102"> one </Word>
            <Word dur="0.25" stime="5.04" conf="0.244"> Christopher </Word>
            <Word dur="0.25" stime="5.04" conf="0.120"> Christian </Word>
            <Word dur="0.10" stime="5.44" conf="0.271"> her </Word>
            <Word dur="0.23" stime="5.58" conf="0.951"> first </Word>
            <Word dur="0.25" stime="5.81" conf="0.969"> came </Word>
            <Word dur="0.10" stime="6.06" conf="0.970"> to </Word>
            <Word dur="0.23" stime="6.16" conf="0.551"> me </Word>
            <Word dur="0.23" stime="6.16" conf="0.195"> meet </Word>
            <Word dur="0.14" stime="6.80" conf="0.913"> if </Word>
            <Word dur="0.20" stime="6.94" conf="0.969"> he </Word>
            <Word dur="0.22" stime="7.14" conf="0.969"> had </Word>
            <Word dur="0.04" stime="7.36" conf="0.937"> a </Word>
            <Word dur="0.37" stime="7.40" conf="0.762"> grip </Word>
            <Word dur="0.15" stime="8.07" conf="0.923"> and </Word>
            <Word dur="0.08" stime="8.26" conf="0.521"> a </Word>
            <Word dur="0.08" stime="8.26" conf="0.188"> the </Word>
            <Word dur="0.23" stime="8.34" conf="0.518"> good </Word>
            <Word dur="0.23" stime="8.34" conf="0.142"> grid </Word>
```

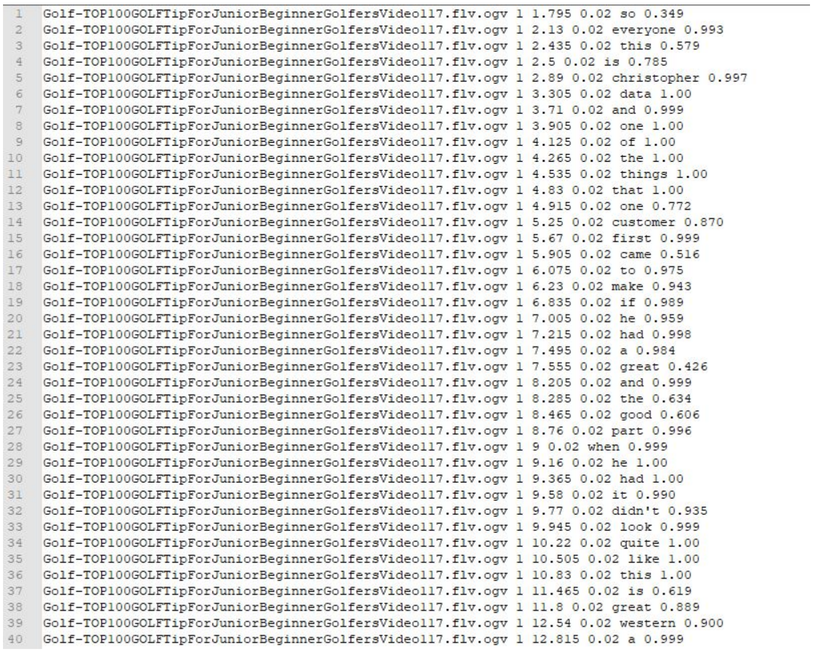
Fig. 2.5 A sample of LIMSI 2016 transcript

concepts. Specifically, 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 for Stopword elimination procedure [42].

2.3.3 Indexing

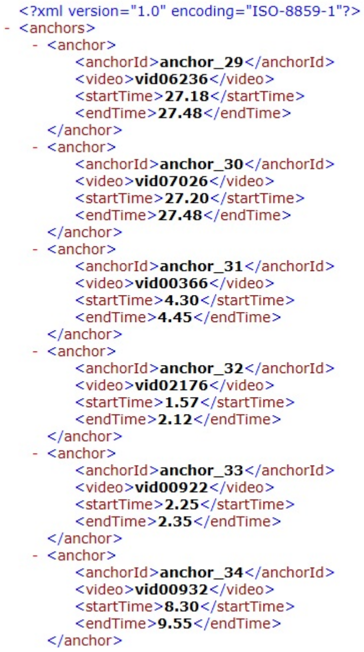
In order to find relevant video segments, we used Apache Solr[1][43] version 6.6 to index the textual and visual features associated with each segment. Figure 2.11 shows the graphical web interface of administrator for the Apache Solr 6.6. Multiple indexes have been created for the video segments, each based on one of the following features: (i) the LIMSI transcripts of the segments, (ii) the visual concepts of the

[1] http://lucene.apache.org/solr



```
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 1.795 0.02 so 0.349
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 2.13 0.02 everyone 0.993
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 2.435 0.02 this 0.579
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 2.5 0.02 is 0.785
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 2.89 0.02 christopher 0.997
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 3.305 0.02 data 1.00
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 3.71 0.02 and 0.999
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 3.905 0.02 one 1.00
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 4.125 0.02 of 1.00
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 4.265 0.02 the 1.00
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 4.535 0.02 things 1.00
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 4.83 0.02 that 1.00
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 4.915 0.02 one 0.772
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 5.25 0.02 customer 0.870
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 5.67 0.02 first 0.999
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 5.905 0.02 came 0.516
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 6.075 0.02 to 0.975
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 6.23 0.02 make 0.943
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 6.835 0.02 if 0.989
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 7.005 0.02 he 0.959
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 7.215 0.02 had 0.998
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 7.495 0.02 a 0.984
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 7.555 0.02 great 0.426
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 8.205 0.02 and 0.999
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 8.285 0.02 the 0.634
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 8.465 0.02 good 0.606
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 8.76 0.02 part 0.996
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 9 0.02 when 0.999
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 9.16 0.02 he 1.00
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 9.365 0.02 had 1.00
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 9.58 0.02 it 0.990
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 9.77 0.02 didn't 0.935
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 9.945 0.02 look 0.999
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 10.22 0.02 quite 1.00
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 10.505 0.02 like 1.00
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 10.83 0.02 this 1.00
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 11.465 0.02 is 0.619
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 11.8 0.02 great 0.889
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 12.54 0.02 western 0.900
Golf-TOP100GOLFTipForJuniorBeginnerGolfersVideo117.flv.ogv 1 12.815 0.02 a 0.999
```

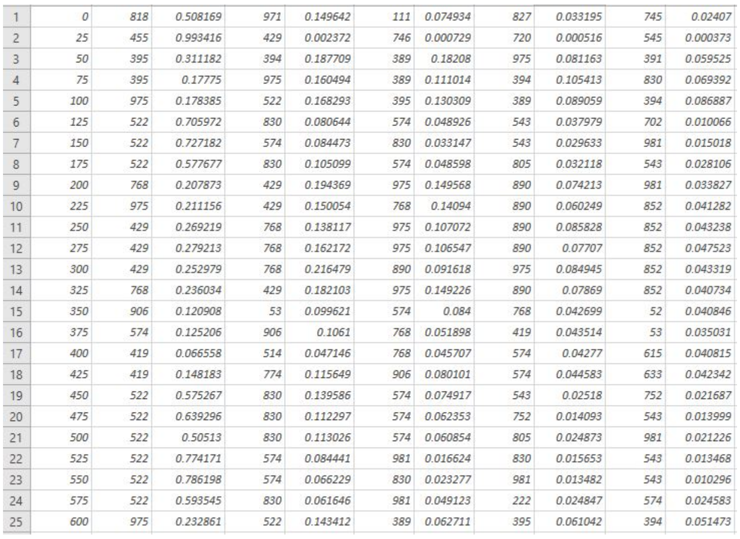
Fig. 2.6 A sample of LIUM 2012 transcript



```
  <?xml version="1.0" encoding="ISO-8859-1"?>
- <anchors>
  - <anchor>
        <anchorId>anchor_29</anchorId>
        <video>vid06236</video>
        <startTime>27.18</startTime>
        <endTime>27.48</endTime>
    </anchor>
  - <anchor>
        <anchorId>anchor_30</anchorId>
        <video>vid07026</video>
        <startTime>27.20</startTime>
        <endTime>27.48</endTime>
    </anchor>
  - <anchor>
        <anchorId>anchor_31</anchorId>
        <video>vid00366</video>
        <startTime>4.30</startTime>
        <endTime>4.45</endTime>
    </anchor>
  - <anchor>
        <anchorId>anchor_32</anchorId>
        <video>vid02176</video>
        <startTime>1.57</startTime>
        <endTime>2.12</endTime>
    </anchor>
  - <anchor>
        <anchorId>anchor_33</anchorId>
        <video>vid00922</video>
        <startTime>2.25</startTime>
        <endTime>2.35</endTime>
    </anchor>
  - <anchor>
        <anchorId>anchor_34</anchorId>
        <video>vid00932</video>
        <startTime>8.30</startTime>
        <endTime>9.55</endTime>
    </anchor>
```

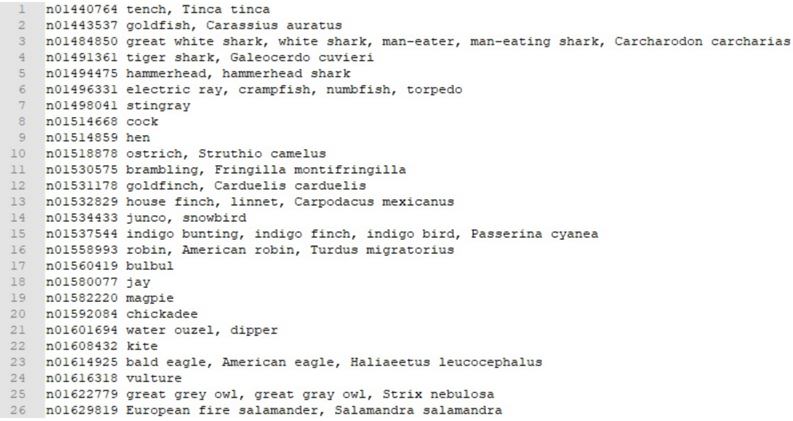
Fig. 2.7 A sample of training anchors of TRECVID 2017

segments (for this feature, we consider only the name of each concepts identified by the visual concept annotation step), and (iii) the metadata of the full videos.



1	0	818	0.508169	971	0.149642	111	0.074934	827	0.033195	745	0.02407
2	25	455	0.993416	429	0.002372	746	0.000729	720	0.000516	545	0.000373
3	50	395	0.311182	394	0.187709	389	0.18208	975	0.081163	391	0.059525
4	75	395	0.17775	975	0.160494	389	0.111014	394	0.105413	830	0.069392
5	100	975	0.178385	522	0.168293	395	0.130309	389	0.089059	394	0.086887
6	125	522	0.705972	830	0.080644	574	0.048926	543	0.037979	702	0.010066
7	150	522	0.727182	574	0.084473	830	0.033147	543	0.029633	981	0.015018
8	175	522	0.577677	830	0.105099	574	0.048598	805	0.032118	543	0.028106
9	200	768	0.207873	429	0.194369	975	0.149568	890	0.074213	981	0.033827
10	225	975	0.211156	429	0.150054	768	0.14094	890	0.060249	852	0.041282
11	250	429	0.269219	768	0.138117	975	0.107072	890	0.085828	852	0.043238
12	275	429	0.279213	768	0.162172	975	0.106547	890	0.07707	852	0.047523
13	300	429	0.252979	768	0.216479	890	0.091618	975	0.084945	852	0.043319
14	325	768	0.236034	429	0.182103	975	0.149226	890	0.07869	852	0.040734
15	350	906	0.120908	53	0.099621	574	0.084	768	0.042699	52	0.040846
16	375	574	0.125206	906	0.1061	768	0.051898	419	0.043514	53	0.035031
17	400	419	0.066558	514	0.047146	768	0.045707	574	0.04277	615	0.040815
18	425	419	0.148183	774	0.115649	906	0.080101	574	0.044583	633	0.042342
19	450	522	0.575267	830	0.139586	574	0.074917	543	0.02518	752	0.021687
20	475	522	0.639296	830	0.112297	574	0.062353	752	0.014093	543	0.013999
21	500	522	0.50513	830	0.113026	574	0.060854	805	0.024873	981	0.021226
22	525	522	0.774171	574	0.084441	981	0.016624	830	0.015653	543	0.013468
23	550	522	0.786198	574	0.066229	830	0.023277	981	0.013482	543	0.010296
24	575	522	0.593545	830	0.061646	981	0.049123	222	0.024847	574	0.024583
25	600	975	0.232861	522	0.143412	389	0.062711	395	0.061042	394	0.051473

Fig. 2.8 A sample of visual concept that is in CSV format



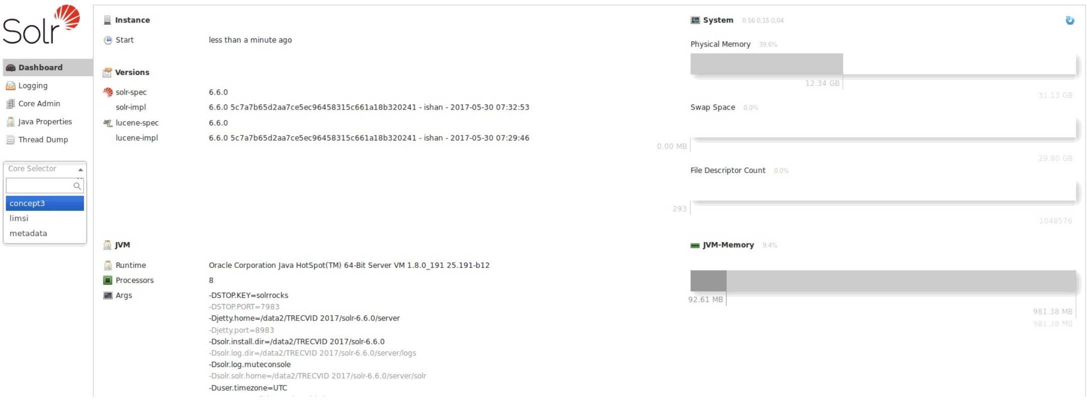
```
n01440764 tench, Tinca tinca
n01443537 goldfish, Carassius auratus
n01484850 great white shark, white shark, man-eater, man-eating shark, Carcharodon carcharias
n01491361 tiger shark, Galeocerdo cuvieri
n01494475 hammerhead, hammerhead shark
n01496331 electric ray, crampfish, numbfish, torpedo
n01498041 stingray
n01514668 cock
n01514859 hen
n01518878 ostrich, Struthio camelus
n01530575 brambling, Fringilla montifringilla
n01531178 goldfinch, Carduelis carduelis
n01532829 house finch, linnet, Carpodacus mexicanus
n01534433 junco, snowbird
n01537544 indigo bunting, indigo finch, indigo bird, Passerina cyanea
n01558993 robin, American robin, Turdus migratorius
n01560419 bulbul
n01580077 jay
n01582220 magpie
n01592084 chickadee
n01601694 water ouzel, dipper
n01608432 kite
n01614925 bald eagle, American eagle, Haliaeetus leucocephalus
n01616318 vulture
n01622779 great grey owl, great gray owl, Strix nebulosa
n01629819 European fire salamander, Salamandra salamandra
```

Fig. 2.9 A sample of synset words for visual concepts

The specific indexing structure implemented by Solr is known as inverted index. An inverted index stores, for each term, the list of documents in which the term is present. This makes term-based queries very efficient [44], and it is exploited by the proposed approach.

The transcripts exploited by our proposed approach are provided by the LIMSI tool, as in our experiments on the training anchors, on average the LIMSI [5] transcripts allow to achieve better results than the LIUM [4] ones.

Fig. 2.10 A sample of a video keyframe related to the previous data samples



```
Instance
Start           less than a minute ago

Versions
solr-spec       6.6.0
solr-impl       6.6.0 5c7a7b65d2aa7ce5ec96458315c661a18b320241 - ishan - 2017-05-30 07:32:53
lucene-spec     6.6.0
lucene-impl     6.6.0 5c7a7b65d2aa7ce5ec96458315c661a18b320241 - ishan - 2017-05-30 07:29:46

JVM
Runtime         Oracle Corporation Java HotSpot(TM) 64-Bit Server VM 1.8.0_191 25.191-b12
Processors      8
Args            -DSTOP.KEY=solrrocks
                -DSTOP.PORT=7983
                -Djetty.home=/data2/TRECVID 2017/solr-6.6.0/server
                -Djetty.port=8983
                -Dsolr.install.dir=/data2/TRECVID 2017/solr-6.6.0
                -Dsolr.log.dir=/data2/TRECVID 2017/solr-6.6.0/server/logs
                -Dsolr.log.muteconsole
                -Dsolr.solr.home=/data2/TRECVID 2017/solr-6.6.0/server/solr
                -Duser.timezone=UTC
```

Fig. 2.11 Apache Solr web interface

2.3.4 Query formulation and segment retrieval

In this stage, we first transform the anchor (query) segment into a textual query by including in the text of the query all the textual information associated with the anchor (i.e., the LIMSI transcripts and the relevant visual concepts) and also the meta-data of the video containing the anchor (i.e., title and tags of the video containing the anchor).

Named-entity recognition is applied on LIMSI to extract the important names inside the query and give them a higher relevance. Named Entity Recognition (NER) labels sequences of words in a text which are related to the names of things, such as person and company names, or gene and protein names.

Concept mapping technique, which is based one WordNet, is used to find the most relevant concepts inside the query. 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 for both concepts list and Metadata, 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.[2]

After the query preparation phase, a tool executes it by using Apache Solr and returns the related segments ranked by relevance.

2.4 Mono-modal Query formulation

In the proposed system, we considered four different mono-modal queries. for each feature, a set of monomodal queries executes and afterwards, they are combined in order to execute multimodal queries and retrieve more relevant video segments. These monomodal queries are used as building blocks of the tested solutions. The specific mono-modal queries are described in the current Section, whereas their combinations are presented in Section 2.5.

Name-entity recognition 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. Name-entity recognition never used alone as a monomoidal query and it is always combined with another feature like LIMSI transcripts.

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. For this reason, each visual concept (its “name”) 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.

The metadata information can be used to: (i) select segments, (ii) select videos. The metadata information is available only at the video level (the same metadata for all the segments of a video). For this reason the metadata information can be used to build a textual query, but it must be executed on the transcripts of the segments if we are interested in selecting segments (this is the only approach we can use to

[2] https://oxforddictionaries.com

select segments by using metadata). The main problem of this approach is that all the anchors contained in the same video will be associated with the same textual query and hence the query will select the same segments, independently of the anchor (this is true only for the anchors contained in the same video). Differently, when metadata are used to retrieve entire videos, we can execute the query on the metadata.

The characteristics of the four monomodal queries are the followings:

1. **LIMSI-based query + Named-Entity Recognition.**

   For each anchor, a textual query is built by considering the words appearing in the LIMSI transcript of the anchor. Then, Named-Entity Recognition (NER) is applied on the anchor LIMSI transcripts to extract relevant names of entities and give them higher relevance in the query. NER labels sequences of words in a text which are related to the names of entities, for instance people and company names, or gene and protein names. The basic idea is that the segments containing the same entities as the anchor are potentially more relevant, hence a higher weight is assigned to those words in the query, as well as groups of 2, 3, and 4 adjacent words, e.g., "United States of America".

   The resulting query is executed on the LIMSI transcript index.

   For example, if the LIMSI text is: *"Handmade portraits: Staceyrebecca"*, the query would be: *"Handmade portraits" (W1.0) OR "Staceyrebecca" (W1.6)* since "Staceyrebecca" is a know entity and it is assigned a higher weight (1.6 instead of 1 in our case, the parameter value of query boost weight is discussed in Section 2.6.2).

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 (the parameter value of visual concept filter is discussed in Section 2.6.2).

   Furthermore, a concept mapping technique based on WordNet is applied to find the most relevant concepts inside the query. The mapping is performed by using the words appearing in the full video metadata and the list of visual concepts of the segment. To maximize the word-list enrichment for concepts and metadata, we applied WordNet using both the synonyms and hypernyms of the words. Furthermore, also groups of 2, 3 and 4 adjacent words are considered.

The concept mapping technique aims at increasing the relevance of the visual concepts of the considered anchor that are related to the content of the whole video. For this reason, each visual concept name of the anchor is compared with the words appearing in the metadata of the video containing the anchor. If the visual concept, or its synonym (or hypernym) based on WordNet, appears in the metadata of the video, then the visual concept is assigned a higher weight in the query. The resulting query is executed on the visual concept index.

For example, if metadata text is: *"Top 100 golf tips for kids"*, and visual concepts are: *"digital clock, golf ball"*, the resulting query would be: *"digital clock" (W1.0) OR "golf ball" (W1.6)*, since *"golf"* is matching.

3. Metadata-based query for segment selection.

Metadata can be used to select either segments or videos. Metadata are associated to the full video, i.e., all segments of a video share the same metadata. A textual query built from a segment (anchor) metadata will be the same for all segments of the same video.

If the query is executed on a metadata index, only full videos can be selected, with all their corresponding segments. Instead, to select specific segments, metadata queries are executed on the LIMSI transcript index, since transcripts are specific for each segment. Named-Entity Recognition (NER) is applied to extract relevant entities and give them higher relevance in the query, by following the same procedure described for queries #1 and #2.

For example, if metadata is: *"United Kingdom weekly Talk Show"*, the query on LIMSI transcripts would be: *"United Kingdom" (W1.6) OR "weekly" (W1) OR "Talk Show" (W1.6).*

4. Metadata-based query for video selection.

This query is the same as the previous one, but it is executed on the metadata index, hence returning videos and not segments. For this reason, the results of such query cannot be used directly to propose the resulting segments, since all the segments of the related videos would be selected. However, this query helps in filtering a pre-selection of videos among which related segments are highly likely to be found (see Section 2.5.2).

2.5 Query combinations for segment retrieval

For the Video Hyperlinking, we designed four different approaches. The considered approaches use different features and/or combine them by using different strategies. Before selecting the configurations of the four approaches, we performed a set of experiments on the training anchors to evaluate the impacts of the two available transcript tools (LIUM vs LIMSI [4, 5]) and two video segmentation techniques (shot segmentation vs fixed length segmentation). On the average, on the training anchors, the LIMSI transcripts allow achieving better results than the LIUM ones and Fixed-segmentation allows retrieving more relevant segments than the shot segmentation-based approach. Hence, the four approaches use the LIMSI transcripts and fixed-segmentation (120 seconds).

In the previous year, we considered a system which used the multimodal queries [41]. But this kind of implementation increased the potential of noises and we got also a very low precison. Based on this experience, For this year, our system is working in multiModal but it uses monomodal queries. Then based on the idea of how we use these monomodal queries, we designed 3 multimodal algorithms:

1. Automatic Feature Selection (AFS) (Section 2.5.1)
2. Metadata-based approach (Section 2.5.2)
3. Pipeline approach (Section 2.5.3)
4. LIMSI-NER approach (Section 2.5.4)

We also consider one monomodal approach that is LIMSI-NER approach (2.5.4), because this approach is embedded in the other approaches and for this reason, we decided to evaluated the effect of this approach separately.

For each of the four combinations, an experimental run has been submitted to TRECVID, and its results are presented in Section 2.6.

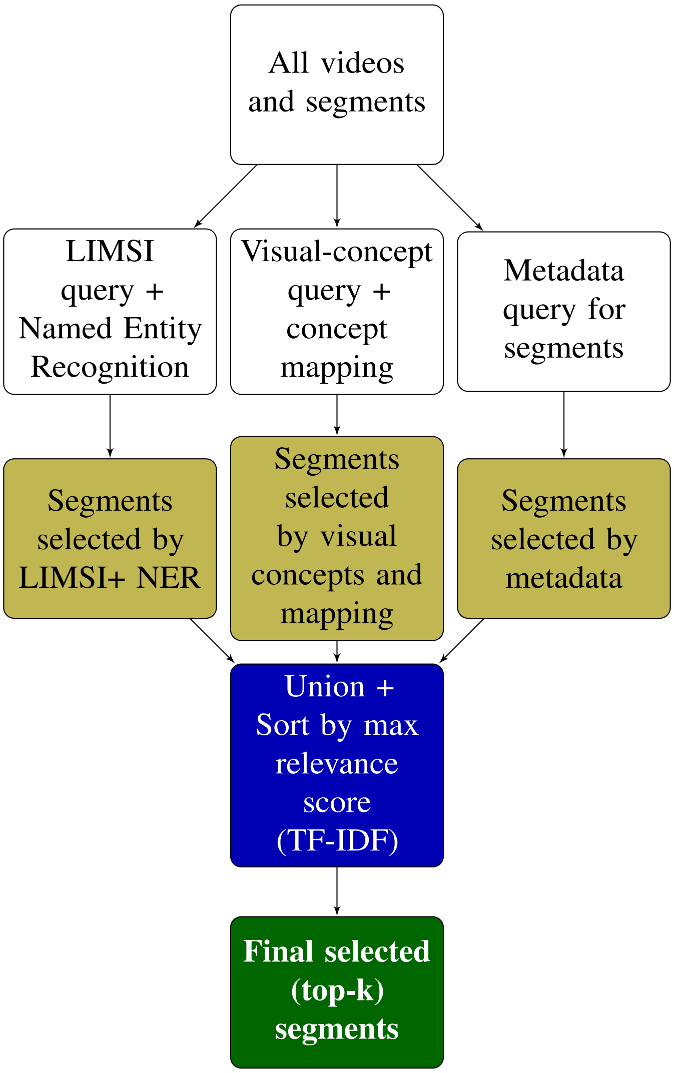

All videos and segments
LIMSI query + Named Entity Recognition
Visual-concept query + concept mapping
Metadata query for segments
Segments selected by LIMSI+ NER
Segments selected by visual concepts and mapping
Segments selected by metadata
Union + Sort by max relevance score (TF-IDF)
Final selected (top-k) segments

Fig. 2.12 Automatic Feature Selection (AFS)

2.5.1 Automaticle Feature Selection (AFS)

In the AFS approach, we used the following features: Meta-data (which are available only at the level of videos), the LIMSI transcripts and Visual concepts. We also applied a Named-entity recognition (NER) technique to identify entities in the textual queries generated for each anchor and a Concept mapping technique to identify the visual concepts, of each anchor, that are semantically related to the metadata of the video of which the anchor is part of. During the execution of the query, a higher importance is given to the words associated with the entities identified by NER and the visual concepts that are selected by the concept mapping technique. This run exploits all the available features and all the building blocks/components of our system. The main idea of AFS approach is to dynamically selects the best feature for each query.

The AFS approach is based on two steps. In the first step, AFS considers one feature at a time and selects the subset of relevant segments for each feature by means of monomodal queries (one monomodal query for each feature). For each returned segment, Solar returns also a relevance score. If a segment is returned twice, the system keeps only a copy of the segment and selects the highest score value among the ones returned by the monomodal queries. In the second step, the subsets of segments retrieved in the first step are merged[3] and ranked in terms of relevance score. The output of this second step is the final result of this approach. Hence, a segment is ranked high, in the final set of returned segments, if it has been associated with a high relevant score with respect to at least one feature. (see Figure 2.12)

2.5.2 Metadata-based approach

Similarly to AFS approach, also this second run uses all the components of our system. Specifically, it considers all the features and also the named-entity recognition (NER) and the concept mapping techniques. However, differently from AFS approach, Metadata are used to perform an initial filter on the videos that could contain interesting segments. In the initial filtering step, for each anchor, the Meta-data based approach selects the videos that are similar to the video containing the anchor under consideration. This video selection is based on the value of the meta-data, which are available at the video level. The basic idea for this approach is that only the segments inside relevant videos should be of interest.

In the second step, the Metadata based approach selects the most relevant segments from the selected videos by using the same approach used in AFS approach. However, only LIMSI and visual concepts are considered in this second step (see Figure 2.13).

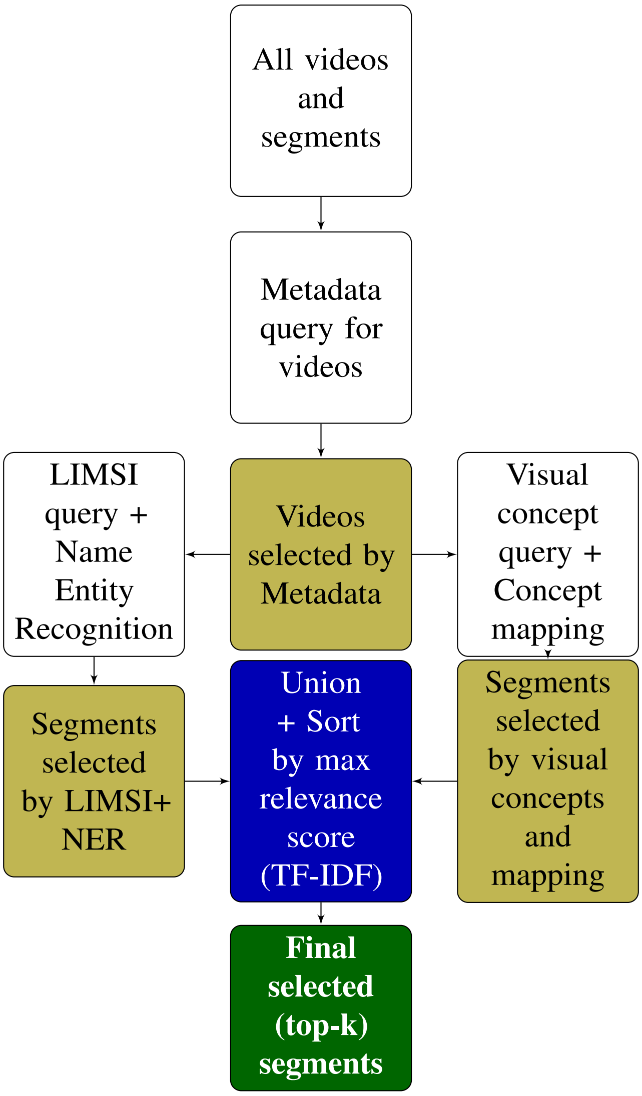

All videos and segments
Metadata query for videos
LIMSI query + Name Entity Recognition
Videos selected by Metadata
Visual concept query + Concept mapping
Segments selected by LIMSI+ NER
Union + Sort by max relevance score (TF-IDF)
Segments selected by visual concepts and mapping
Final selected (top-k) segments

Fig. 2.13 Metadata based approach

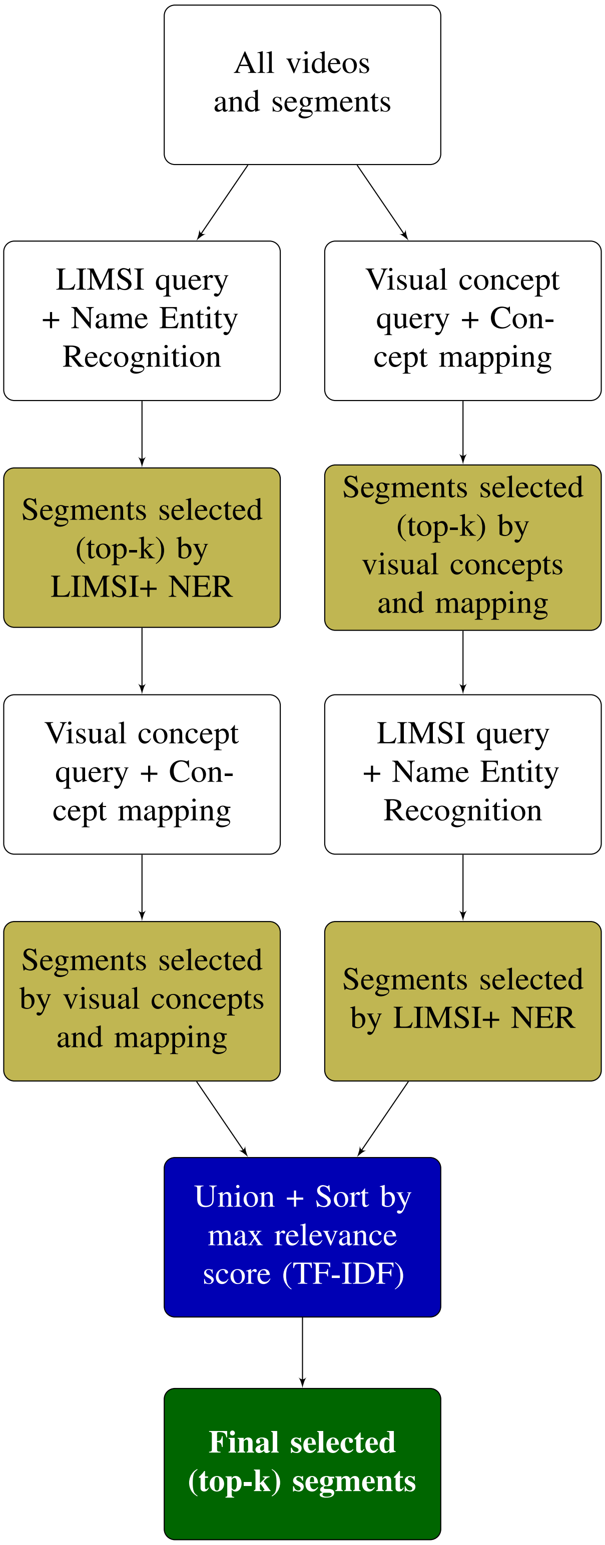

All videos and segments
LIMSI query + Name Entity Recognition
Visual concept query + Concept mapping
Segments selected (top-k) by LIMSI+ NER
Segments selected (top-k) by visual concepts and mapping
Visual concept query + Concept mapping
LIMSI query + Name Entity Recognition
Segments selected by visual concepts and mapping
Segments selected by LIMSI+ NER
Union + Sort by max relevance score (TF-IDF)
Final selected (top-k) segments

Fig. 2.14 Pipeline approach

2.5.3 Pipeline approach

For this approach, we used only two features: LIMSI and Visual concepts. Also in this run, we applied the Named-entity recognition (NER) and the Concept mapping techniques. This approach aims at selecting the segments that are relevant with respect to both features. In fact, a segment is selected if and only if it is selected by both features singularly. The main goal for implementing this approach was to use and analyze only machine generated data at the segment level.

In the Pipeline approach, for each anchor, we first select the top-k relevant segments by using a query based on LIMSI and then we refine the result by querying the subset of returned segments by means of a query based on the visual concept feature. The same operation is then performed by switching the order of the two queries. Finally, the two subsets of returned segments are merged and ranked in terms of relevance score. (see Figure 2.14) The sequence ***LIMSI + Visual Concepts*** is not equal to ***Visual Concepts + LIMSI*** because only the top-k segments are selected in the first step.

In parallel to this approach, we proposed another approach which was similar to the first step of Pipeline approach. It used LIMIS for the first step and then takes into account the visual concepts. We could not submit this run at TRECVid 2017, because we could only submit four runs. So as th results were not officially evaluated by TRECVID, we discarded this approach. Although, the results on training sets in compare to the results of Pipeline approach, were lower in terms of both Precision and MAiSP.

2.5.4 LIMSI-NER approach

In this approach, we considered only the LIMSI transcript feature and we applied the named-entity recognition (NER) technique on the queries. The aim of this approach is to analyze the differences between Monomodal and Multimodal techniques. We selected LIMSI for the monomodal approach because it achieved the best results on the development anchors (it is better than the Monomodal approach based on

[3] Note that duplicate segments are removed and the highest score is selected among the copies of each segment.

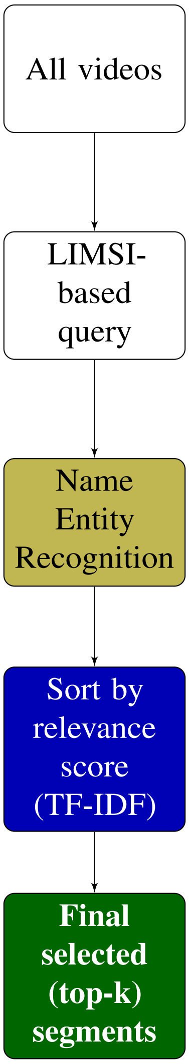

All videos
LIMSI-based query
Name Entity Recognition
Sort by relevance score (TF-IDF)
Final selected (top-k) segments

Fig. 2.15 LIMSI-NER approach

visual concept and also better than the one based on the LIUM transcripts). (see Figure 2.15)

2.6 Experimental evaluation

The four proposed approaches have been submitted to the TRECVID 2017 video hyperlinking benchmark and their results are presented in Section 2.6.1, whereas the impact of the set of parameter values are discussed in Section 2.6.2.

Metric	AFS	Pipeline	LIMSI-NER	Metadata
P@5	**0.840**	0.808	0.725	0.704
P@10	**0.808**	0.748	0.667	0.556
MAP	**0.164**	0.114	0.093	0.082
MAiSP	**0.253**	0.185	0.155	0.132

Table 2.1 Results of the different approaches submitted to TRECVID according to each evaluation metric.

2.6.1 Experimental results

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), which considers true positives all segments overlapping with a segment that was considered relevant in the ground truth [45].
- An adapted MAP called Mean Average interpolated Segment Precision (MAiSP) [36]

Table 2.1 reports the results provided by TRECVID 2017 for each of our approaches.

AFS (Automatic Feature Selection) approach yields the best results in term of all the considered metrics. We recall that it exploits all the available features (LIMSI transcripts, visual concepts, and Metadata) by executing three monomodal queries (one for each feature) and then merge the selected segments and rank them in terms of relevance score. The chart 2.16 shows the total number of occurrence of each modality for each anchor and for the top 10 segments. Visual concepts and Metadata are the most features that are selected in the top 10 segments. Table 2.2 shows the impact of each modality by describing the percentage of average performance per modality over all the anchors. For example, when Metadata is selected, 89.1% of selected segments are related. However, only 75.2% of selected segments by Visual concepts are related, although the total number of selected segments by visual concepts are the most segments appeared in top 10. Based on this analysis, in AFS

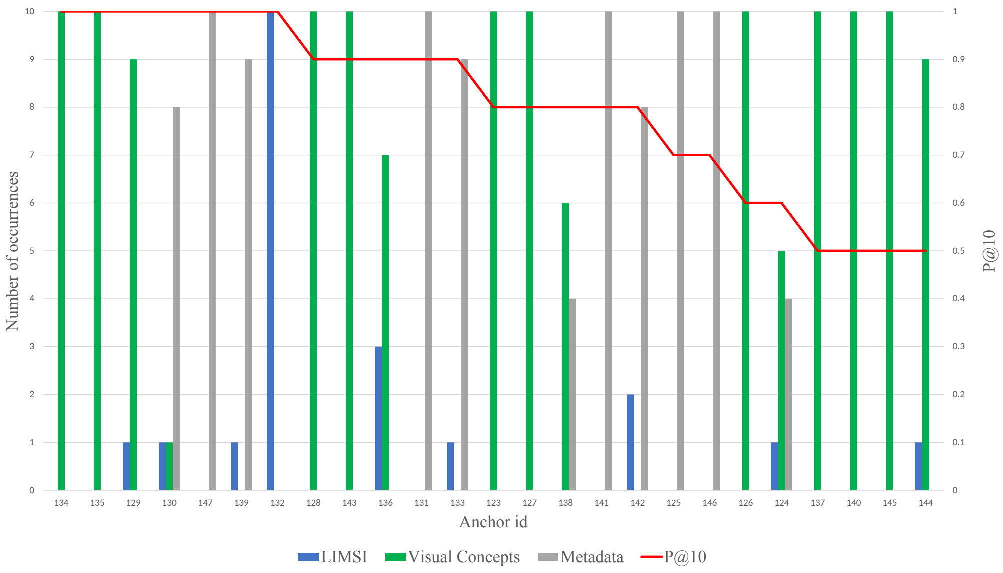


Fig. 2.16 Composition of AFS results: for each test anchor, the number of relevant segments provided by each query (LIMSI, visual concepts, and video metadata) within the top 10 resulting segments are reported, together with the total number of actually relevant segments (P@10).

approach, when Metadata is selected, the selected segments are indeed better than the segments selected by LIMSI or Visual Concepts.

Pipeline approach is characterized by high values for all metrics. However, it performs worse than AFS approach. Hence, pipeline the queries seem to have a negative impact on the final results. Another difference between Pipeline approach and AFS approach is that in Pipeline approach we do not consider the Meta-data feature. Hence, in some cases, it probably allows selecting relevant segments.

Meta-data based approach achieved the lowest result. This was slightly unexpected as performance of this run on the development anchors was higher in compare to those of Pipeline approach and LIMSI-NER approach. This is most likely due to the fact that using the Meta-data for pre-filtering videos would raise the problem of selecting very few related videos for some anchors. Hence, for some anchors this approach returns few segments.

Finally, the results confirm that the exploitation of more features is usually better than using one single feature (the results of LIMSI-NER approach, which is based only on LIMSI, are on the average lower than those of AFS approach and Pipeline

AFS queries	P@10	Number of segments
Metadata	0.891	92
LIMSI	0.762	21
Visual concepts	0.752	137

Table 2.2 Average P@10 of each query contributing to the AFS, over all the test anchors.

approach). Also by comparing the AFS approach and Pipeline approach, the role of Metadata for improving the result could be more visible.

Figure 2.17 reports the comparison of approaches based on MAiSP measure for the three teams that participated in TRECVid 2017. AFS approach (Automatic Feature Selection) ranked first in this competition. Pipeline approach and LIMSI-NER are the ranked in following. However, Metadata based approach ranked 5. In terms of precision at rank 5 (see figure 2.22) and 10 (P@5 and P@10), AFS approach and Pipeline approach ranked after the approaches of the VIREO team. There were not very much difference between the precision of our approaches and the VIREO team's approaches. However, the most important measure for TRECVid is MAiSP as it is based on the start and end of segments and evaluates the whole segment.

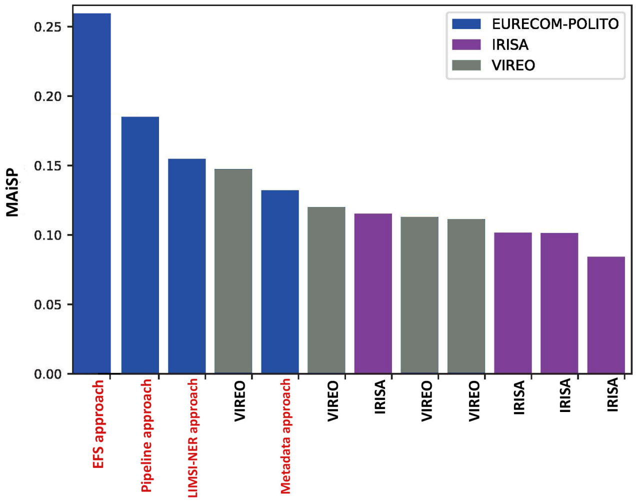


Fig. 2.17 Results of all the approaches submitted to TRECVID 2017 in terms of MAiSP

Measures	**AFS**	Metadata	Pipeline	LIMSI-NER
P@10	**0.289**	0.227	0.221	0.212
MAP	**0.096**	0.077	0.071	0.065
MAiSP	**0.084**	0.062	0.059	0.054

Table 2.3 Pre-evaluation results based on ground-truth

2.6.2 Analysis on the impact of parameters

In order to improve the performance of algorithms, an analysis is being done on the impact of parameters for the developed algorithms. The ground truth were provided by the TRECVid organizers and contains the top 10 ranks for the training anchors. However, there a lot of segments which are not annotated by the judgments and in our analysis, we consider these segments as a not related segment. For this reason, the measures P@5, MAiSP or MAP are not demonstrated well the effect of changes on the parameters. So for this analysis, in order to select the best option, we considered only precision at rank 10 (P@10).

The table 2.3 shows the pre-evaluation results on the approaches using ground-truth for training anchors that are provided by TRECVid 2017 Organizers. However, these results contains around 60% not evaluated segments because the ground-truth segments are not covered for all the dataset and only contain a partial evaluated segments which are evaluated on TRECVid 2016. Figure 2.18 describes for each approach over all the 90 training anchors, the mean percentage of tags for the top 10 segments and for all the 90 training anchors, how many of the segments are accepted, rejected or not yet evaluated based on the ground-truth segments.

Results on the training set (Table 2.3) and on the test set (Table 2.1) lead to the same top-performing algorithm, i.e., AFS. However they rank the other approaches differently: metadata, pipeline, and LIMSI in the training set, and pipeline, LIMSI and metadata in the test set. The most noteworthy difference is the metadata approach, which is the second-best on the training set and becomes the worst on the test set. We consider that such data-dependent changes in results are due to the small number of samples in both datasets, so few specific samples can influence the overall ranking of the approaches.

The default parameter-value configuration considered for all approaches is as follows.

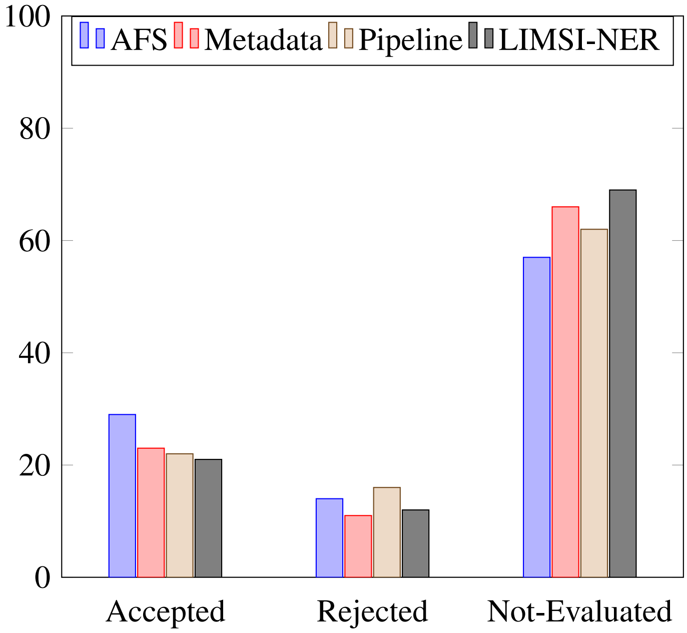


Fig. 2.18 % of pre-evaluation segment tags based on ground-truth for training anchors

- K-filter: **1000**
- Stemming algorithm: **SnowballPorter**
- Filter threshold of visual concepts: **0.3**
- Query boost weight: **1.6**
- NER classifier: **Multi Classifier**
- WordNet similarity algorithm: **Lin**
- Lin algorithm threshold: **0.7**

K-filter indicates the top-k number of segments to be kept in the final step of each approach: it is fixed to 1000 because in TRECVID each participant/approach was allowed to submit up to 1000 segments for each run.

The analysis of the other parameters is described in following.

1. Stemming algorithms in Solr

To be able to search the text efficiently and effectively, Solr splits the text into tokens during both indexing and query execution. Those tokens can also be

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

Table 2.4 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

Table 2.5 P@10 results for filter threshold of visual concepts

pre- and post-filtered for additional flexibility. This allows for case-insensitive search, misspelled product names, synonyms, etc. [46]. For our approaches, we analyzed four stemming token filters:

1. PorterStem transforms the token stream by applying the Porter stemming algorithm.
2. SnowballPorter stems words using a Snowball-generated stemmer.
3. Hunspell is a TokenFilterFactory that creates instances of HunspellStemFilter.
4. KStem is a high-performance kstem filter for English.

Based on the analysis done on our four approaches (see table 2.4), the SnowballPorter is the best stemmer in terms of precision at rank 10.

2. Threshold of visual concept recognition

In order to remove noises and optimize using of visual concepts for the approaches, a set of threshold for filtering the concepts is analyzed (see table 2.5). This parameter is not applied on the LIMSI-NER approach.

Based on this analysis, The threshold 0.5 and 0.7 are removing more noises and they are selecting more accurate concepts. However, using a higher threshold causes of lacking concepts for some anchors and reduced the final precision.

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

Table 2.6 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**

Table 2.7 P@10 results for NER classifiers

For this reason, threshold 0.3 is considered.

3. Query boost weight

Query boost value parameter is used for Concept mapping technique and Named Entity Recognition (NER). The aim of using this parameter is to give a higher weight to the selected query words when the query executes in Solr [47].

To achieve the best boosting value while doing search, a set of parameters analyzed (see table 2.6). Regarding the results achieved in the analysis, the boosting factor 1.6 is the best value in order to obtain the highest precision at rank 10.

4. NER classifier

Stanford NER is also known as CRFClassifier. It provides a general implementation of (arbitrary order) linear chain Conditional Random Field (CRF) sequence models. There are two kinds of CRFs provided by Stanford Named Entity Recognizer: Single CRF NER Classifier and Multiple CRFs NER Classifier. We analyzed these two Classifiers on our approaches (see table 2.7). Based on the analysis, Multiple CRFs NER Classifier ranked first among all

Algorithm	AFS	Metadata	Pipeline	LIMSI-NER
LESK	0.279	0.217	0.209	0.198
Lin	**0.289**	**0.227**	**0.221**	**0.212**
Wu-Palmer	0.281	0.220	0.208	0.202

Table 2.8 P@10 results for WordNet similarity algorithms

Threshold	AFS	Metadata	Pipeline	LIMSI-NER
0.6	0.281	0.218	0.214	0.203
0.7	**0.289**	**0.227**	**0.221**	**0.212**
0.8	0.275	0.215	0.210	0.198

Table 2.9 P@10 results for Lin algorithm threshold

approaches in terms of P@10.

5. WordNet similarity for concept mapping

The aim in using WordNet similarity is to assign a quantitative value to the related words. We considered four different algorithms for this analysis:

1. The Wu-Palmer (Wu & Palmer) [48] calculates relatedness by considering the depths of the two synsets in the WordNet taxonomies, along with the depth of the LCS (Least Common Subsumer).
2. Resnik [49] similarity score denotes how similar two word senses are, based on the Information Content (IC) of the Least Common Subsumer (most specific ancestor node).
3. Lin [50] adapts Resnik's method and defines the similarity of two concepts as the ratio between the amount of information needed to state the commonality between them and the information needed to fully describe them.
4. LESK [51] metric measures the overlap between the glosses of the two concepts and also concepts directly related via relations such as hypernyms and meronyms.

Based on the analysis results (see table 2.8), Lin algorithm is performed better in terms of precision at rank 10 in our approaches.

In order to improve Lin similarity algorithm, a threshold used to filter the selected mapped concepts. Although a previously analysis was done on this threshold [52], but because the dataset of this work is different, we decided to analyze again this threshold for a short range of values. Table 2.9 demonstrates the results achieved for this analysis. The current results confirm the previous study, also indicate that using threshold 0.7 will improve the final precision at rank 10.

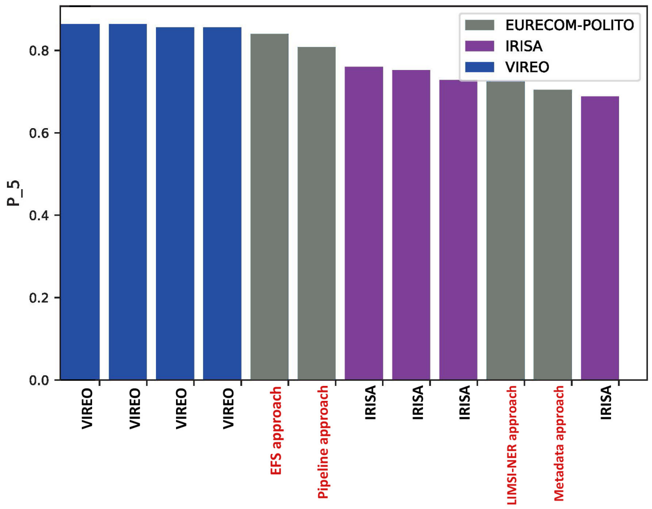


Fig. 2.19 TRECVid 2017 rankings in terms of precision at rank 5 (P@5)

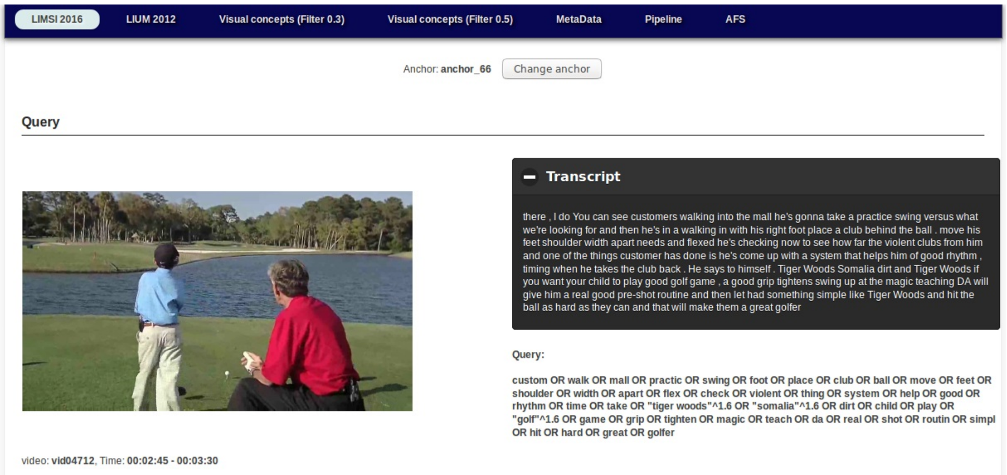
2.6.3 Visualizing of Video hyperlinking

We created a Web interface in order to analyze the effect of each feature on the query and comparing the results with the provided ground-truth.

In the Web interface, it is possible for the users to select different algorithms, also different anchors and then it will display (i) the video (and its information, i.e., Id, duration, etc.), the selected feature data and the generated query for this anchor, and (ii) the result segments associated to the selected anchor and for the selected algorithm. Also it is possible to see the ground-truth videos of the selected anchor and the feature data related to them. The later allows experts to compare the results together and also with ground-truth. Moreover, identify how the ground-truth is being selected and what was the problem of the returned video which were rejected in the ground-truth.

This system interface can be used with 2 different purposes. The first one, if you are developing a new solution and you want to check how the system works, you need something to shows you a certain video and the features of the selected video in order to understand if the system is returning the right segments. Furthermore, it allows you to see how the system works, and also to check if there are some errors in the ground-truth. We decided to use a different color for each of the returned video in order to show if the returned video is a part of the ground-truth, we use green border for that video which shows this segment is relevant. If the segment was considered as not-relevant in the ground-truth, we use a red border and if the segment was not evaluated in the ground-truth, we use an orange border. The other purpose of this system is that we want to use this system also to evaluate the new algorithm by allowing an expert to select a right answer (a right segment) provided by our algorithm.

Figure 2.20 shows a screenshot of a part of interface where the video, the feature information of the video (in this image is LIMSI 2016 transcript) and the generated textual query for the feature. Figure 2.21 demonstrates the part related to comparison of the top selected segments based on the generated query with the segments of the Ground-truth. Furthermore, it is possible to evaluate the information of the selected feature (in this example is LIMSI 2016 transcript) for both Ground-truth and the top selected segments.



Anchor: **anchor_66**

Query

Transcript

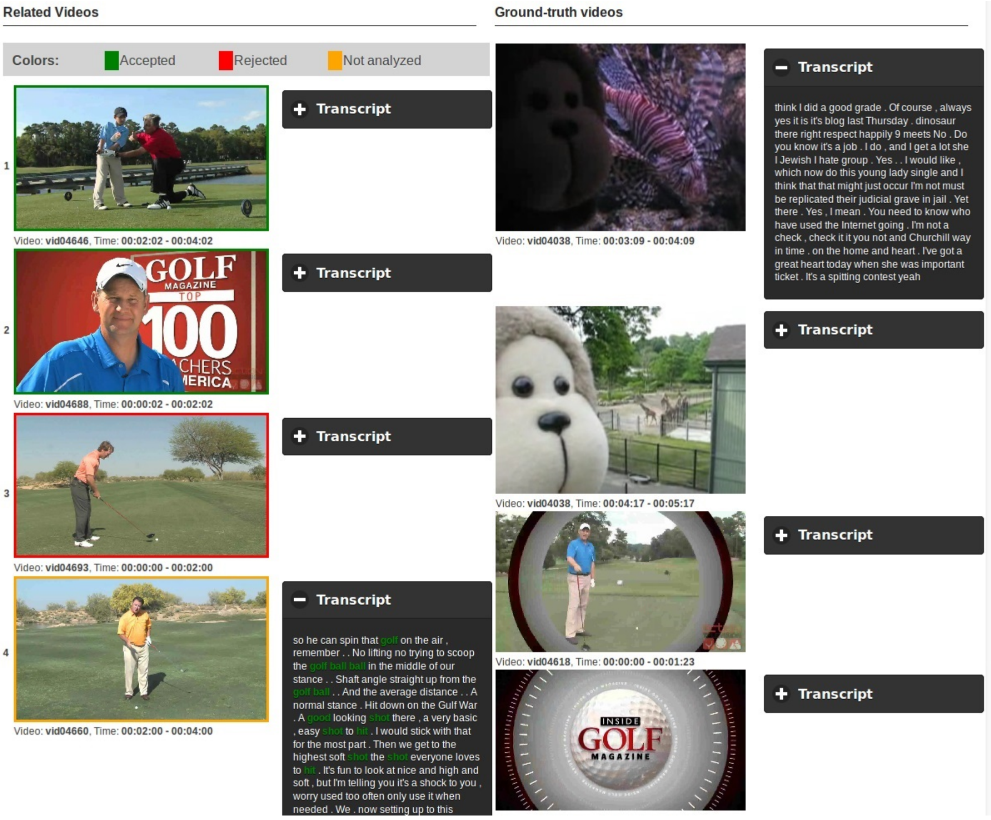
there , I do You can see customers walking into the mall he's gonna take a practice swing versus what we're looking for and then he's in a walking in with his right foot place a club behind the ball . move his feet shoulder width apart needs and flexed he's checking now to see how far the violent clubs 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

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

Fig. 2.20 The query visualization interface.


Related Videos
Ground-truth videos
Colors: 🟩 Accepted 🟥 Rejected 🟧 Not analyzed
1
2
3
4
+ Transcript
− Transcript
Video: vid04646, Time: 00:02:02 - 00:04:02
Video: vid04688, Time: 00:00:02 - 00:02:02
Video: vid04693, Time: 00:00:00 - 00:02:00
Video: vid04660, Time: 00:02:00 - 00:04:00
Video: vid04038, Time: 00:03:09 - 00:04:09
Video: vid04038, Time: 00:04:17 - 00:05:17
Video: vid04618, Time: 00:00:00 - 00:01:23
GOLF MAGAZINE TOP 100
INSIDE GOLF MAGAZINE

so he can spin that golf on the air , remember . . No lifting no trying to scoop the golf ball ball in the middle of our stance . . Shaft angle straight up from the golf ball . . And the average distance . . A normal stance . Hit down on the Gulf War . A good looking shot there , a very basic , easy shot to hit . I would stick with that for the most part . Then we get to the highest soft shot the shot everyone loves to hit . It's fun to look at nice and high and soft , but I'm telling you it's a shock to you , worry used too often only use it when needed . We . now setting up to this

think I did a good grade . Of course , always yes it is it's blog last Thursday . dinosaur there right respect happily 9 meets No . Do you know it's a job . I do , and I get a lot she I Jewish I hate group . Yes . . I would like , which now do this young lady single and I think that that might just occur I'm not must be replicated their judicial grave in jail . Yet there . Yes , I mean . You need to know who have used the Internet going . I'm not a check , check it it you not and Churchill way in time . on the home and heart . I've got a great heart today when she was important ticket . It's a spitting contest yeah

Fig. 2.21 The comparison visualization interface.

2.7 Analysis of Visual concepts

The goal of this section is to evaluate concepts detected by the visual concept detector in order to improve the selection of mapped concepts for our system.

The total number of concepts used by the visual concept detector is 1000. For each keyframe, visual concept detects 50 concepts. For each of these concepts, it reports the concept ID (which we could easily get the name of concept by referring to the synset file of concepts) along with a relevant score. Based on our data segmentation (Section 2.3.2), we have 4 keyframes in average for each segments which means 200 concepts would selected in each segment.

Concerning that using a lot of mapped concepts, would cause loosing the effect of other concepts, we considered only the top 10 mapped concepts (after applying the threshold on concepts). These mapped concepts are selected from the concepts

Concept	Frequency
croquet ball	16
ballplayer	14
baseball player	14
sweatshirt	13
bottlecap	11
lakeside	8
lakeshore	8
park bench	8
parallel bars	6
bars	6
worm fence	4
snake fence	4
snake-rail fence	4
Virginia fence	4

Table 2.10 Frequency of detected concepts

that have the highest score and they are the most frequent concepts among the other concepts.

For the sake of simplicity, we consider one segment to show how this analysis select the concepts. This segment is from the video 4712 and it started from second minute of the video and lasts for 120 seconds. Figure 2.10 shows a sample keyframe of this segment.

For this segment, 200 concepts are detected. However, by removing the duplicate concepts and by applying the filter threshold (0.3) over them, only 14 concepts were remain.

Table 2.10 shows the frequency of the concepts. Frequency is of a concept represents the number of times a concept appears in the segment. Table 2.11 represents the maximum relevant score for each concept. In this plot, if a concept appears more than one time, only the maximum relevant score would be considered.

By comparing both of the figures, it is clear that most of frequent concepts have also the highest score among other concepts. Finally, 10 concepts have been selected for this segment:

croquet ball , ballplayer, baseball player, lakeside, lakeshore , park bench , parallel bars, bars , bottlecap , sweatshirt

Concept	Score
croquet ball	0.703310
ballplayer	0.669513
baseball player	0.639513
lakeside	0.564288
lakeshore	0.554436
park bench	0.516501
bottlecap	0.496688
parallel bars	0.471018
bars	0.471018
sweatshirt	0.385856
worm fence	0.348585
snake fence	0.335425
snake-rail fence	0.326447
Virginia fence	0.319621

Table 2.11 Relevant score of detected concepts

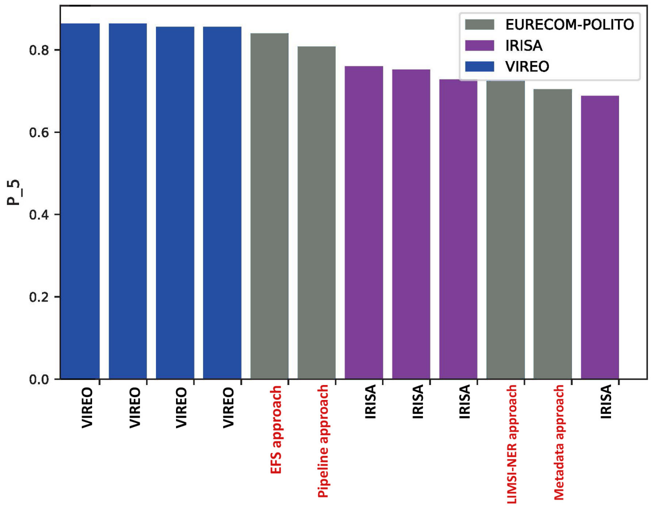


Fig. 2.22 TRECVid 2017 rankings in terms of precision at rank 5 (P@5)

One of the good way for analysis of concepts is analysis based on the score of each visual concept. We can focus on the concept which has the best high scores like scores above 0.95 and analyze them based on the shots that they have high score.

We are looking for the concepts that have high scores in some shots and on the other hand, they have low scores in the other shots. **Distribution of total no. of concepts for different thresholds:** Figure 2.23 shows the analysis regarding the distribution of total number of concepts that are visible in different thresholds. We can see how quickly the number of concepts decreased when the threshold increased. The range of threshold is from 0.1 to 0.95. For the threshold 0.95, we have 14,594 concepts, while for the threshold 0.1, we have 1,539,493 concepts.

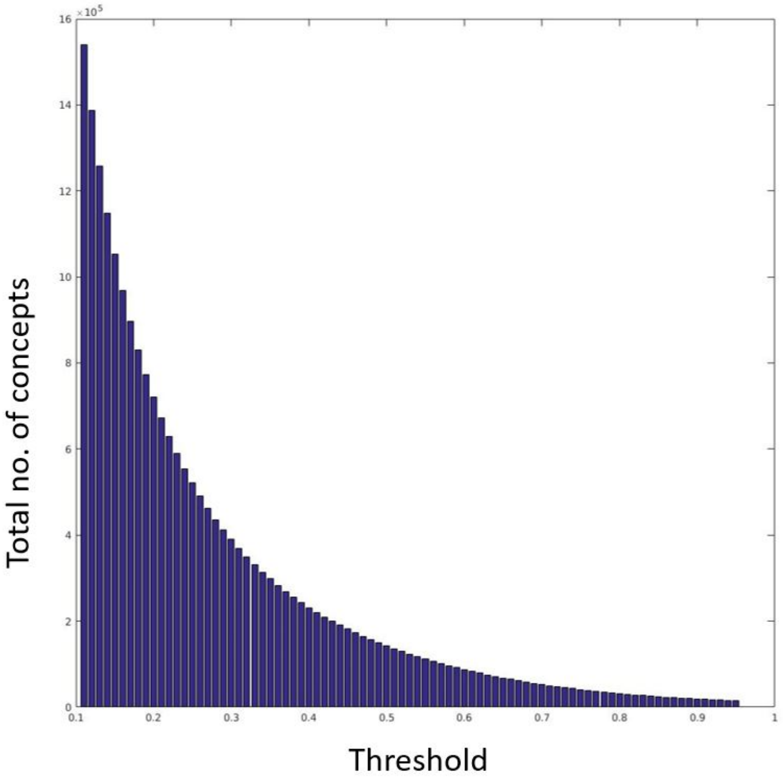


Fig. 2.23 Distribution of total no. of concepts for different thresholds

Maximum of total no. of concepts per shot for different thresholds:

In figure 2.24, the maximum of the total number of concepts per shot for the threshold from 0.1 to 0.95 is analyzed. As it is shown by the plot, there is not a lot of concepts that has been visible if we increase the threshold.

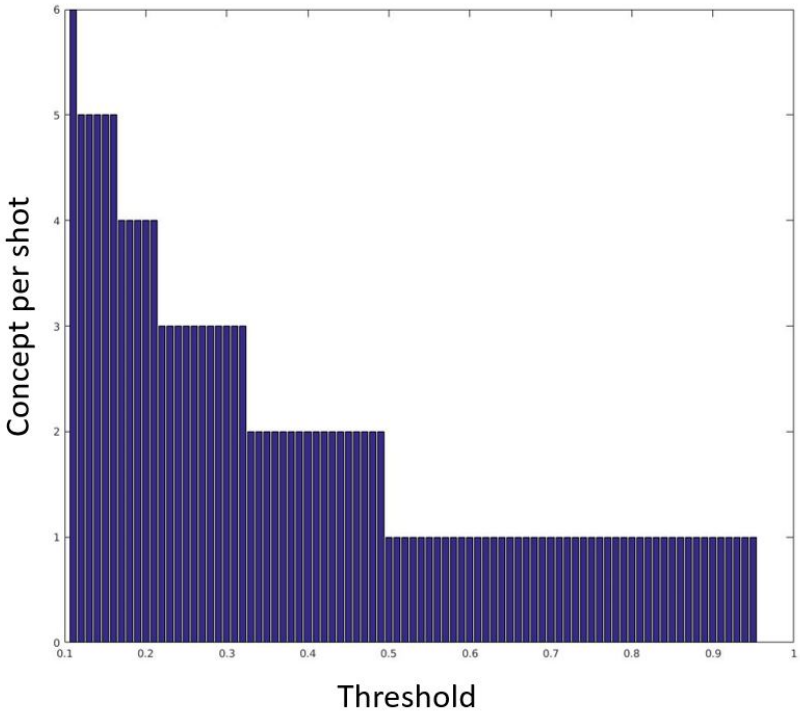


Fig. 2.24 Maximum of total no. of concepts per shot for different thresholds

Chapter 3

Discovering cross-topic collaborations among researchers

In recent years a huge amount of publications and scientific reports has become available through digital libraries and online databases. Digital libraries commonly provide advanced search interfaces, through which researchers can find and explore the most related scientific studies. Even though the publications of a single author can be easily retrieved and explored, understanding how authors have collaborated with each other on specific research topics and to what extent their collaboration have been fruitful is, in general, a challenging task. This thesis proposes a new pattern-based approach to analyzing the correlations among the authors of most influential research studies. To this purpose, it analyzes publication data retrieved from digital libraries and online databases by means of an itemset-based data mining algorithm. It automatically extracts patterns representing the most relevant collaborations among authors on specific research topics. Patterns are evaluated and ranked according to the number of citations received by the corresponding publications. The proposed approach was validated in a real case study, i.e., the analysis of scientific literature on genomics. Specifically, we first analyzed scientific studies on genomics acquired from the OMIM database to discover correlations between authors and genes or genetic disorders. Then, the reliability of the discovered patterns was assessed using the PubMed search engine. The results show that, for the majority of the mined patterns, the most influential (top ranked) studies retrieved by performing author-driven PubMed queries range over the same gene/genetic disorder indicated by the top ranked pattern.

3.1 Related work

This work is partly related to the following research topics: (i) Author-Topic Model, (ii) Graph-based co-authorship models, (iii) Citation content analysis, (iv) Reviewer assignment, and (v) Weighted association rule mining. Hereafter, we will separately overview each topic and discuss the position of our work with respect to existing studies.

Author-Topic Model. The problem of modeling the interests of authors on different topics based on textual document analysis has already been investigated in literature. The Author-Topic Model (ATM) [53] is a generative model for textual documents, where topics are represented as probability distributions over words while authors are associated with probability distributions over topics. The ATM allows us to represent the original documents as a mixture of topics and to determine which authors have mainly contributed to a given topic. For example, given a set of publications the corresponding research topics can be extracted first. Then, the subset of most active researchers on each topic can be extracted. [54] have proposed a Bayesian approach to estimate the ATM parameters. Since the ATM correlates single authors with specific topics, it cannot be directly applied to infer cross-topic collaborations among multiple authors.

Graph-based co-authorship models. Graph- and network-based models have already been adopted to model co-authorship relationships (e.g., [55–58]). Specifically, in [55, 56] graph theory and visualization models have jointly been exploited to model co-authorship and citation relations. [58] used a graph indexing technique (i.e., PageRank [59]) to identify the most authoritative researchers. The relationships among researchers can be also modeled as social networks. For example, ArnetMiner [57] is a social network of academic researchers, where for each author a research profile is automatically extracted from the Web and integrated with publication data accessible through existing digital libraries. Network- and graph-models represent connections between authors without explicitly considering the correlations with the covered topics. Therefore, the underlying information differs from those provided by the patterns considered in this study.

Citation content analysis. To study the impact of scientists' research, the number of citations received by their scientific publications has been considered in several studies (e.g., [60–62]). Citation content analysis is the research branch that focuses on studying citations among papers thus computing a reputation score for each researcher. Specifically, it focuses on analyzing the semantics, syntax, and position in the text of the paper of the citations to reveal the influence of both authors and scientific papers. For example, [61] analyzed the sentences including citation expressions to identify interesting characteristics of scholarly communication. [60] and [62] classified citations based on their semantics to gain insights into the relationships between authors and topics. In our work, citations are exploited to weigh the relevance of a publication thus, indirectly, to measure the reputation of a group of researchers related to a given topic. However, our analysis is not focused on citation analysis. As discussed in Section 3.2, to measure the relevance of a publication different measures can be easily integrated as well.

Reviewer assignment. A related branch of research concerns the assignments of reviewers to scientific papers. The aim is to support editors in the peer review of scientific papers by automatically recommending potential reviewers. For example, [63–65] addressed the problem of choosing a pool of reviewers for a given paper based on the expertise of a potentially large set of candidate reviewers and on the main topics covered by the paper under review. The authors tackled the optimization problem to assign each paper to at least three independent reviewers with complementary expertise so that the pool of reviewers assigned to each paper covers most of the topics addressed by the paper and each reviewer has a reasonable number of reviews to do. Unlike the works proposed by [63–65] the task addressed in this work is not an optimization problem. The techniques adopted in our work are exploratory and allow us to discover interpretable patterns useful for supporting a number of advanced analyses.

Weighted association rule mining. A parallel research effort has been devoted to efficiently extracting itemsets and association rules from weighted data [66–68, 9]. This problem extends the traditional association rule mining task, which was first introduced by [69] in the context of market basket analysis, to the case in which data items are no longer considered as equally relevant within the analyzed data. For example, in the context of market basket analysis the goal is to find sets of products frequently purchased together by taking into account not only the list of products that customers have put into their market basket but also the purchased amount and unitary price of each purchased product. [9] proposed to extract weighted association rules, i.e., rule including weights denoting item significance are extracted. [68] and [66] used weights to drive the frequent and infrequent itemset mining processes, respectively, while [67] automatically generated weights by means of graph indexing techniques. This work focuses on extracting weighted association rules from publication data to discover cross-topic collaborations among authors. A preliminary version of this work has been presented by [70]. This work extends its preliminary version to a large extent. The main differences can be summarized as follows:

(i) Topics are characterized as probability distributions over words which are automatically extracted from publication documents and not only selected from publication metadata.

(ii) Weighted Association Rules (WARs), which represent implications between combinations of authors and topics, are extracted as well on top of frequent itemsets. The newly extracted patterns measure the strength of an implication between authors and topics (e.g., to what extent the citations received by a group of researchers are related to a specific topic) and not only the observed frequency of appearance of a combination of authors in the publication dataset. To the best of our knowledge, this work is the first attempt to exploit WARs to analyze cross-topic collaborations among authors.

(iii) WARs are classified based on their goal into five main categories. WAR categories allow us to identify not only topic-specific collaborations but also *cross-topic* collaborations among authors.

3.2 Cross-topic Scientific collaboration analyzer

Cross-topic Scientific Collaboration Analyzer (CSCA) is a data-driven methodology to automatically discover significant cross-topic collaborations among authors of scientific publications. The methodology is based on the application of an exploratory data mining technique on the publication data which are retrieved from digital libraries or online databases such as PubMed [7] and OMIM [8].

The goal is to identify groups of co-authors who have significantly contributed to the research community related to a particular topic or to a given set of topics. The relevance of the scientific productions of a group is the number of citations that is received by the co-authored publications.

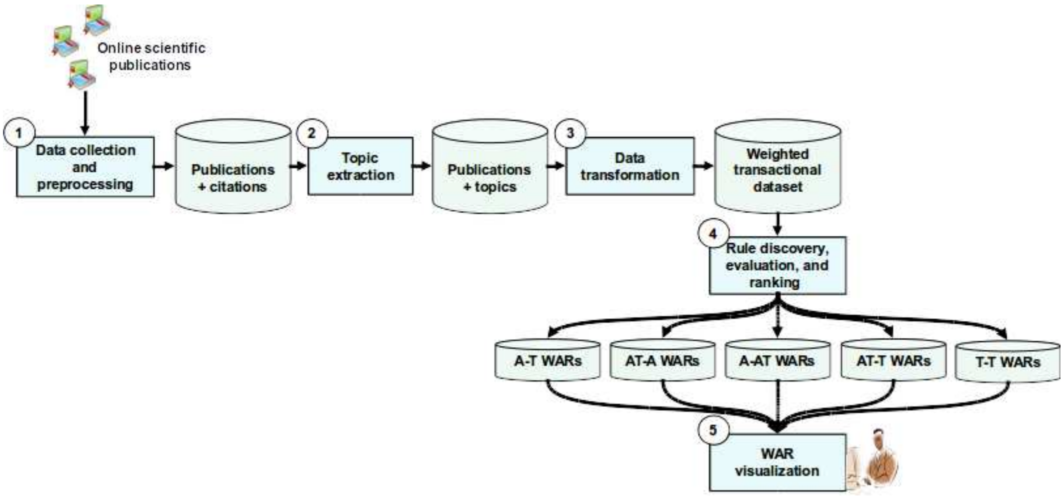


Fig. 3.1 Cross-topic Scientific Collaboration Analyzer

For each identified group of authors CSCA extracts, classifies, and ranks patterns, called Weighted Association Rules (WARs), that allow us to answer to the following questions:

(1) Which is the topic that the collaboration among researchers focused on?

(2) How many topics are considered in the collaboration?

(3) By considering each topic separately, What is the relevance of their scientific production?

(4) On which topics the group of authors collaborated with external authors?

(5) To what extent are the topics addressed in the collaborations correlated with each other?

Figure 3.1 describe the methodology which consists of five main steps:

(i) *Data collection and preprocessing.* Publications data and related metadata are retrieved from online sources, preprocessed in order to prepare them for the next mining process, and then stored into a centralized repository (see Section 3.2.1).

(ii) *Topic extraction.* The topics of each publication are gained from either publication metadata or from the textual content of the publication by using the Author-Topic Model (see Section 3.2.2).

(iii) *Data transformation.* Author information, citation counts, and publication topics are prepared to the association rule mining step (see Section 3.2.3).

(iv) *Rule discovery, evaluation, and ranking.* Weighted Association Rules (WARs), which represent implications between combinations of authors and topics, are extracted, classified, and ranked to support knowledge discovery from publication data (see Section 3.2.4).

(v) *Rule visualization.* The mined WARs are visualized through a Web-based application in order to allow the exploring the results more easily. (see Section 3.2.5).

In following, each step is described in more detail.

3.2.1 Data collection and preprocessing

Publication data are retrieved from digital libraries and online databases (e.g., PubMed [7], OMIM [8]) by using the Application Programming Interfaces (APIs) that provided by the used sources and then interpreted and finally stored in a unique repository.

For our purposes, for each publication we acquire the following data:

(i) the Digital Object Identifier (DOI) of the publication,

(ii) the list of authors,

(iii) the current number of citations received,

(iv) the content text of the publication, and

(v) any metadata that are 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 [71]. Hereafter, we will consider it as reference indicator of the influence/popularity of a publication.

Publication data can be enriched by considering metadata describing the addressed topics. For example, the OMIM database [8] collects publications about genomics and genetics, and for each publication the list of related genes and genetic disorders are given. As discussed in Section 3.2.2, we will consider such information (if available) to identify the main topics covered by each publication.

Before proceeding to the next mining stpes, two established text preprocessing steps are applied to the text of the publications and the related metadata specifically, similarly to what we did in the TRECVID task (Section 2.3), we applied the stopword elimination and the stemming preprocessing technique. To perform stopword elimination, in our experiments we used the Natural Language Toolkit (NLTK) stopword corpus [72].

Furthermore, the author names and the descriptors of genes and genetic disorders are made uniform by removing noisy characters, abbreviated forms, etc.

3.2.2 Topic extraction

A list of covered topics is assigned to each publication. Depending on the data source, topics can be retrieved from metadata (e.g., genes and genetic disorders in the OMIM database [8]) or be unknown (e.g., PubMed [7]). We propose two strategies for assigning topics to each publication. (i) if topic are given in metadata, CSCA used metadata content as descriptors of the topic. (ii) otherwise, by using the Author-Topic Model (ATM) [53], CSCA will extract a description of the main topic of each publication from its content.

ATM is a generative model for textual documents, where documents in the input collection are modeled as mixture of topics. Each topic is represented as probability distribution over word stems as described in the Latent Dirichlet Allocation (LDA) model [73]. More specifically, for each publication document a distribution over topics is first sampled from a Dirichlet distribution. Next, for each word stem in the document a single topic is assigned according to the distribution. Finally, each word stem is sampled from a multinomial distribution over word stems specific to the sampled topic [53]. In the computation, the generative algorithm keeps track of a $W \times T$ (word stem-by-topic) and a $A \times T$ (author-by-topic) count matrices. The algorithm starts by assigning word stems to random topics and authors from the set of authors and documents. Count matrices are stored from 10 samples (with random initial assignments) at the 2000th iteration of the Gibbs sampler. From the count matrices topics and authors are extracted. Each topic is characterized by (i) word-based description W_{de}, i.e., the top-10 word stems that are most likely to be generated conditioned on the topic, and (ii) author-based description A_{de}, i.e., the top 10 most likely authors to have generated a word stem conditioned on the topic.

For each publication document, we extract the top-k main topics by following the procedure described in Algorithm 3.1. For finding the word stems, we scan the input document that are included in the description W_{de} of any topic in T. For each topic, we store the maximum per-word count in $W \times T$ over all words in its description. Since word counts indicate the relevance of word in the topic, we assign the top-k topics associated with the word stems with maximal count.

Algorithm 3.1: Main topic detection

```
Require: the publication documents $D$, the word stem-by-topic count matrix $W \times T$, and the word stem descriptions $W_{de}$ of
    all topics $T$
Ensure: set of main topics $t^* \in T$ for each document $d$ in $D$
for all $d$ in $D$ do
    $top[t]$=0 $\forall$ $t \in T$
    for all word stem $w$ occurring in $D$ do
        for all topic $t$ in $T$ do
            if $w \in W_{de}$ then
                update $top[t]$ if the $w$'s count in $W \times T$ is higher than the current $top[t]$ value
            end if
        end for
    end for
    select the top-k topics of $d$ associated with the k maximal values in $top$
end for
return the top-K topics of each document $d$
```

3.2.3 Data transformation for WAR mining

In order to min WARs, publication data, citation scores, and topics are stored into a weighted transactional dataset. A weighted transactional dataset is a set of pairs $\langle$transaction, weight$\rangle$, where each *transaction* corresponds to a different scientific publication, while *weight* is the value of the citation counter of the represented publication (see Section 3.2.1).

Transactions consist of sets of items, where items are publication authors (e.g., *Smith, L.*), or research topics (e.g., *topic X*). Topics can be described either by the metadata content or by the ATM description (see Section 3.2.2). Items are represented in the form (*feature*:*value*), where *feature* is *Author* or *Topic*, while *value* is the corresponding feature value.

A more formal definition of weighted transactional dataset is given below.

Definition 1 Weighted transactional dataset. *Let A be the set of authors and T be the set of topics. Let P be the set of all scientific publications and let $C(p_i)$ ($p_i \in P$) be an influence score associated with publication p_i. An item i_k is a pair feature:v_q, where $v_q \in A$ if feature is equal to Author or $v_q \in T$ if feature is equal to Topic. A transaction t_j is a set of items related to publication p_j. A weighted*

Pub. id	Citation count	Authors	Topics
1	10	$(Author{:}Brown,\ J.)$, $(Author{:}Smith,\ L.)$	$(Topic{:}A)$, $(Topic{:}X)$, $(Topic{:}Z)$
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	$(Author{:}Smith,\ L.)$	$(Topic{:}Z)$

Table 3.1 Example of weighted transactional dataset

transactional dataset $\mathscr{D}$ is a set of weighted transactions, where each weighted transaction $tw_j \in \mathscr{D}$ corresponds to a different publication $p_j \in P$ and it consists of a pair $\langle t_j, C(p_j)\rangle$.

For example, Table 3.1 reports an example of dataset consisting of six weighted transactions, each one corresponding to a different scientific publication. Each publication, identified by the respective id, is weighted by the corresponding number of citations (see Column *Citation count*). For each publication the list of authors (see Column *Authors*) and the covered topics (see Column *Topics*) are known. Publications can be co-authored, and can be related to many topics. For example, publication with pub. id 1 received 10 citations (i.e., transaction weight equal to 10). Its corresponding transaction consists of the following items: $Author{:}Brown,\ J.$, $Author{:}Smith,\ L.$, $Topic{:}A$, $Topic{:}X$ and $Topic{:}Z$. The transaction refers to a publication that was co-authored by Brown J. and Smith L. and that relates to topics A, X and Z. This transaction is represented this in format:

$$<\{(Author{:}Brown,\ J.), (Author{:}Smith,\ L.)\}\ , 10> \rightarrow (Topic{:}A), (Topic{:}X), (Topic{:}Z)$$

3.2.4 Pattern discovery, evaluation, and ranking

In this step, an exploratory data mining approach, i.e., the Weighted Association Rule (WAR) mining technique, is applied to the prepared weighted transactional dataset. The goal is to automatically generate patterns, i.e., the WARs, representing interesting implications between combinations of authors and topics. After that, WARs are classified based on their semantic meaning into three main categories and ranked to simplify the manual exploration of the mining result.

This section is organized as follow. Sub-section 3.2.4.1 introduces the concept of WAR and its quality indices, Sub-section 3.2.4.2 provides a high-level description of the algorithm used to extract the WARs of interest. Finally, Sub-section 3.2.4.2 introduces the WAR categories and discusses how they can be exploited to help experts to answer to the research questions introduced at the beginning of this Section 3.2.

3.2.4.1 Weighted association rules

Association rule mining [69] is an established data mining technique for discovering recurrent correlations among data items that are hidden in large datasets. Association rule mining is performed as a two-step process which contains (i) frequent itemset mining from the transactional data and (ii) association rule discovery from the set of frequent itemsets mined at the previous step.

Frequent itemset mining. A k-itemset is a set of k distinct items in a transactional dataset. It points out the co-occurrence of the correlate items in the analyzed dataset. In this analysis, an item represents either an author or a topic (see Definition 1). Hence, itemsets may represent co-occurrences of multiple authors and topics in the analyzed dataset. A more formal definition of itemset is given below.

Definition 2 Itemset. *Let $\mathscr{D}$ be a weighted transactional dataset and let $\mathscr{I}$ be the set of distinct items in the form feature:v_q contained in any weighted transaction $tw_j \in \mathscr{D}$. A k-itemset (i.e., an itemset of length k) is a set of k distinct items in $\mathscr{I}$.*

Note that each itemset may contain an arbitrary number of items belonging to any feature.

Commonly, in itemset mining, a minimum support threshold is considered because generating all the possible itemsets is computationally unfeasible even on medium-size datasets [69]. Given the minsup threshold, the frequent itemset mining extracts all the itemsets that *frequently* occur in the source dataset $\mathscr{D}$, i.e., all itemsets whose frequency of occurrence (support) in the source dataset is above a given threshold *minsup*. The support threshold prevents the extraction of less relevant or misleading itemsets. However, it allows us to consider only the most recurrent and thus potentially reliable patterns.

For example, itemset $\{(Author{:}Brown,\ J.),(Topic{:}X)\}$ is occurred three times in the dataset in Table 3.1 (publications with ids 1, 2, and 5). Therefore, by considering a minimum support threshold *minsup*=2 the itemset would be extracted because its frequency of occurrence (3) is above the minimum (user-provided) threshold.

Unfortunately, the number of frequent itemsets can be very large. To prevent the generation of redundant patterns and to simplify the manual inspection of the result, a more compact subset of frequent itemsets, called the closed itemsets [74], can be used. An itemset I is *closed* if there exists no superset that has the same support of I.

Itemset evaluation based on weighted support. The support quality index of an itemset does not consider the relative importance of each transaction in the source dataset [69]. Moreover, in our context of analysis, each publication may have a different impact on the research community. In order to evaluate pattern significance, pattern occurrence in each publication is weighted according to its impact on the research community. For instance, an itemset occurring in a publication with 0 citations should be weighted more than one occurring in a publication with 1 citation.

As our goal is to generate only the combinations of authors and topics that have gained a high impact, we extended the standard itemset mining problem by integrating item weights [9]. Specifically, item occurrences within each transaction (publication) are weighted by an influence score, such as the citation count (see Section 3.2.1). Therefore, the co-authorship of publications with a large number of citations is rewarded, whereas co-authorship of publications with few citations are penalized. To formalize this step, we introduce the concept of *weighted support* of an itemset as a weighted frequency of occurrence of the itemset in the weighted transactional dataset.

Definition 3 Weighted support of an itemset. *Let $\mathscr{D}$ be a weighted transactional dataset and I be an itemset. Let tw_j: $\langle t_j, C(p_j)\rangle$ be an arbitrary weighted transaction in $\mathscr{D}$. The weighted support of I in $\mathscr{D}$, hereafter denoted by wsup(I), is defined as follows:*

$$wsup(I) = \sum_{tw_j \in \mathscr{D} | I \subseteq t_j} C(p_j)$$

Recalling the previous example, $\{(Author{:}Brown,\ J.),(Author{:}Smith,\ L.),\ (Topic{:}X)\}$ has a weighted support equal to 25 because it covers the weighted transactions with publication ids 1 (weight 10), 2 (weight 5), and 5 (weight 10), respectively.

Weighted association rule discovery. Weighted Association Rules (WARs) are extracted on top of frequent itemsets. Given two itemsets A and B (of arbitrary length) a weighted association rule $A \to B$ is an implication between A and B. A more formal definition follows.

Definition 4 Weighted association rule. *Let A and B be two itemsets. A weighted association rule is represented in the form $R : A \to B$, where A and B are the body and the head of the rule respectively.*

A and B are also denoted as antecedent and consequent of rule $A \to B$. Association rule extraction is commonly driven by weighted support (wsup) and confidence (wconf) quality indexes [69]. While the weighted support index represents the weighted frequency of occurrence of the rule in the source dataset, the weighted confidence index represents the rule strength.

Definition 5 Weighted support of a WAR. *Let $\mathscr{D}$ be a weighted transactional dataset. The weighted support (wsup) of a weighted association rule $R : A \to B$ is defined as the weighted support of $A \cup B$ in $\mathscr{D}$.*

Definition 6 Weighted confidence of a WAR. *Let $\mathscr{D}$ be a weighted transactional dataset. The weighted confidence (wconf) of a weighted association rule $R : A \to B$ is the conditional probability of (weighted) occurrence in $\mathscr{D}$ of itemset B given itemset A, i..e,*

$$wconf(R) = \frac{wsup(R)}{wsup(A)} = \frac{wsup(A \cup B)}{wsup(A)}$$

.

For example, WAR $\{(Author{:}Brown,\ J.),(Author{:}Smith,\ L.)\} \to (Topic : X)\}$ shows an implication between a couple of authors and a specific topic. The WAR has weighted support equal to 25 and weighted confidence equal to $\frac{25}{35}$ (= 71,43%), because the implication holds for publications with ids 1, 2, and 5 but not for publication with id 3 (citation count = 10).

WAR categories. For our purposes, we consider five main categories of WARs.Each category consists of the set of all WARs characterized by a predefined sequence of items (authors and/or topics). Categories are tailored to different research questions.

Category 1: Authors-Topic Rules. These rules are extracted to answer to the following questions: *On what topics is the collaboration focused on? Is the collaboration focused on a specific topic or spread over multiple topics?*

WARs of category Authors-Topic (hereafter denoted as A-T WARs) are represented in the form $R : A \rightarrow B$, where the rule antecedent A is an arbitrary itemset consisting of a set of authors, while the consequent B is an arbitrary itemset including a single topic.

For example, $\{(Author{:}Brown,\ J.),(Author{:}Smith,\ L.)\} \rightarrow (Topic : X)\}$ is an A-T WAR. It indicates that authors J. Brown and L. Smith have co-authored publications related to topic X. $\{(Author{:}Brown,\ J.),(Author{:}Smith,\ L.)\} \rightarrow (Topic : Z)\}$ is another A-T WAR with the same antecedent, which indicates that the same authors have collaborated on topic Z. If both WARs are extracted, then the two authors have fruitfully collaborated on multiple topics in separate publications. Notice that if the same publications cover multiple topics, part of co-authored publications may cover both topics. We will separately consider this particular case in the WAR category 4 (see *AuthorsTopics-Topic Rules*).

The weighted support of an A-T WAR indicates the sum of the citation counts of all the publications co-authored by the authors approving in the antecedent of the rule. Sorting rules by decreasing wsup allow experts to consider first the research collaborations that have received a fairly high attention from the research community. Notice that WARs with low wsup are early pruned during the mining process (due to support threshold enforcement), because the corresponding collaborations were very unlikely to produce significant results.

The weighted confidence indicates the fraction of citations received by the co-authored publications on the considered topic with respect to the total number of citations received by all the co-authored publications (independently of the topic). Sorting rules by decreasing wconf allows experts to select, among all the topics covered during the collaborations, the topics that have achieved the highest impact for each group of co-authors. A-T WARs with high wconf indicate the topics on which the collaboration is mainly focused on.

Given a combination of authors, the wsup index allows experts to filter out the less relevant collaborations. On the other hand, the wconf value indicates the strength of the correlation between the set of authors and a particular topic. For example, if the wconf of an A-T WAR is close to 100% (all the citations are associated with a particular topic) then it means that the collaborations of the referred co=authors were productive only on the corresponding topic.

Category 2: AuthorsTopic-Author Rules. These rules are extracted because they allow us to answer to the following question: *Working on a given set of topics, has the group (occasionally) collaborated with external authors?*

WARs of category AuthorsTopic-Author (hereafter denoted as AT-A WARs) are represented in the form $R : A \to B$, where the rule antecedent A is an arbitrary itemset consisting of a set of authors and a set of topics, while the consequent B is an arbitrary itemset including a single author.

For example, WAR $\{(Author{:}Brown,\ J.),(Author{:}Smith,\ L.),\ (Topic{:}X)\} \to (Author{:}Black,\ J.)$ is an AT-A WAR. It indicates that in the collaboration between authors J. Brown and L. Smith on topic X they have collaborated with author J. Black. WAR $\{(Author{:}Brown,\ J.),\ (Topic{:}X)\ (Topic{:}Z)\} \to (Author{:}Smith,\ L.)$ is another AT-A WAR which indicates a cross-topic collaboration between a couple of authors.

The weighted support of an AT-A WAR 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. 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 author referred in the rule consequent. Therefore, low wconf value indicate occasional (yet potentially fruitful) collaborations, whereas high wconf values indicate more systematic collaborations between group of co-authors and external authors.

Category 3: Authors-AuthorTopic Rules. These rules are extracted because they allow us to answer to the following question: *Has the group collaborated with external authors? On which topics?*

WARs of category Authors-AuthorTopic (hereafter denoted as A-AT WARs) are represented in the form $R : A \to B$, where the rule antecedent A is an arbitrary itemset consisting of a set of authors, while the consequent B is an arbitrary itemset including a single author and a single topic.

For example, $\{(Author{:}Brown,\ J.),(Author{:}Smith,\ L.)\} \to \{(Author{:}Black,\ J.),\ (Topic : X)\}$ is an A-AT WAR. It 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 support of the AT-A WAR indicates the significance of the collaboration between the group of authors and the consider pair author-topic. 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.

Category 4: AuthorsTopics-Topic Rules. These rules are extracted because they allow us to answer to the following question: *Given a group of researchers who have frequently collaborated on a set of topics, which other topic is likely to be covered by their co-authored publications?*

WARs of category AuthorsTopics-Topic (hereafter denoted as AT-T WARs) 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, $\{(Author{:}Brown,\ J.), (Author{:}Smith,\ L.), (Topic:X)\} \rightarrow \{(Topic:Z)\}$ is an AT-T WAR. It 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. Hence, these rules allow us to measure the strength of the cross-topic authors' collaborations.

Category 5: Topics-Topic Rules. These rules are extracted because they allow us to answer to the following question: *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, an example of Topics-Topic WARs (hereafter denoted as T-T WARs) is $\{(Topic:A), (Topic:X)\} \rightarrow \{(Topic:Z)\}$.

Sorting T-T WARs by decreasing confidence allows us to identify the sets of most correlated sets of topics.

3.2.4.2 The extraction algorithm

Many frequent Weighted Association Rule (WAR) mining algorithms have already been proposed in literature (e.g., [66–68, 9]). To accomplish the WAR mining task from weighted transactional data, we applied to a two-step mining process that requires (i) Closed itemset mining, and (ii) WAR generation from closed itemsets. Step (i) is accomplished by an FP-Growth-based algorithm [75]. The algorithm relies on an FP-tree data model, i.e., a compact, tree-based representation of the original dataset residing in main memory. Itemset extraction is optimized to generate only closed itemsets. Step (ii) focuses on generating WARs from closed itemsets by generating any combinations of closed itemsets representing WARs of interest [76].

3.2.5 WAR visualization

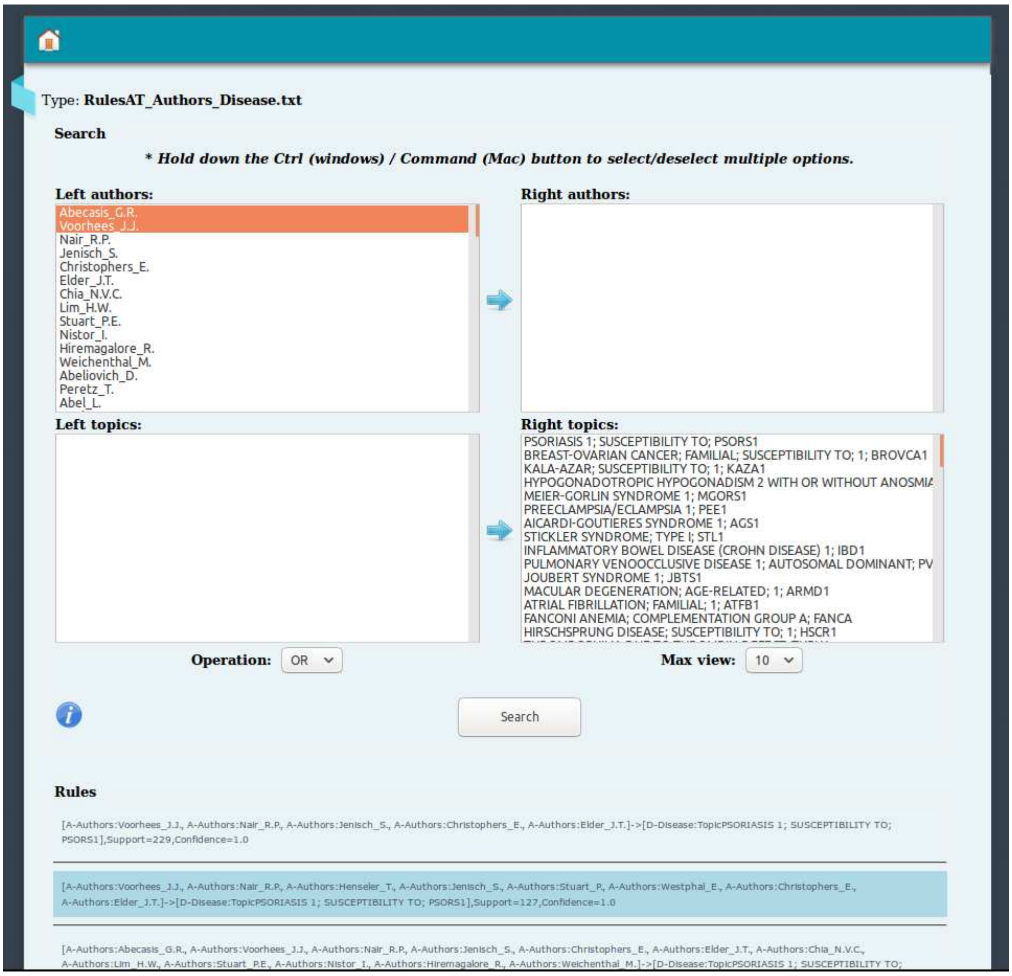
We created a Web interface in order to allow domain experts to browse the rules associate with the category of interest, to filter WARs not including any specific combinations of authors or topics, and to sort the extracted WARs by decreasing weighted confidence.

By using the web interface, after selecting a category, the user can filter among WARs by selecting (i) optional single or multiple authors of the left or the right of WARs, and (ii) optional single or multiple topics on the left or right of WARs. Moreover, The user can select the filter operation type which could be AND/OR and applies on the selected combinations of authors, topics or both. Finally the list of filtered WARs will be displayed which allows easier identifying of which author-topic combinations are potentially of interest for advanced analysis.

Figure 3.2 shows a screenshot of the developed interface. The interface can be accessed at following link: http://dbdmg.polito.it/CSCA/.

3.3 Experimental results

The proposed methodology has been studied in a real case study, i.e., the analysis of the research collaboration on genomics or genetics. Experiments have been performed on the publication data and citations retrieved from the Online Mendelian Inheritance in Man (OMIM) catalog of genetic disorders [8]. The goal in this analysis



Type: **RulesAT_Authors_Disease.txt**

Search

** Hold down the Ctrl (windows) / Command (Mac) button to select/deselect multiple options.*

Left authors:

Abecasis_G.R.
Voorhees_J.J.
Nair_R.P.
Jenisch_S.
Christophers_E.
Elder_J.T.
Chia_N.V.C.
Lim_H.W.
Stuart_P.E.
Nistor_I.
Hiremagalore_R.
Weichenthal_M.
Abeliovich_D.
Peretz_T.
Abel_L.

Right authors:

Left topics:

Right topics:

PSORIASIS 1; SUSCEPTIBILITY TO; PSORS1
BREAST-OVARIAN CANCER; FAMILIAL; SUSCEPTIBILITY TO; 1; BROVCA1
KALA-AZAR; SUSCEPTIBILITY TO; 1; KAZA1
HYPOGONADOTROPIC HYPOGONADISM 2 WITH OR WITHOUT ANOSMIA
MEIER-GORLIN SYNDROME 1; MGORS1
PREECLAMPSIA/ECLAMPSIA 1; PEE1
AICARDI-GOUTIERES SYNDROME 1; AGS1
STICKLER SYNDROME; TYPE I; STL1
INFLAMMATORY BOWEL DISEASE (CROHN DISEASE) 1; IBD1
PULMONARY VENOOCCLUSIVE DISEASE 1; AUTOSOMAL DOMINANT; PV
JOUBERT SYNDROME 1; JBTS1
MACULAR DEGENERATION; AGE-RELATED; 1; ARMD1
ATRIAL FIBRILLATION; FAMILIAL; 1; ATFB1
FANCONI ANEMIA; COMPLEMENTATION GROUP A; FANCA
HIRSCHSPRUNG DISEASE; SUSCEPTIBILITY TO; 1; HSCR1

Operation: OR

Max view: 10

Rules

[A-Authors:Voorhees_J.J., A-Authors:Nair_R.P., A-Authors:Jenisch_S., A-Authors:Christophers_E., A-Authors:Elder_J.T.]->[D-Disease:TopicPSORIASIS 1; SUSCEPTIBILITY TO; PSORS1],Support=229,Confidence=1.0

[A-Authors:Voorhees_J.J., A-Authors:Nair_R.P., A-Authors:Henseler_T., A-Authors:Jenisch_S., A-Authors:Stuart_P., A-Authors:Westphal_E., A-Authors:Christophers_E., A-Authors:Elder_J.T.]->[D-Disease:TopicPSORIASIS 1; SUSCEPTIBILITY TO; PSORS1],Support=127,Confidence=1.0

[A-Authors:Abecasis_G.R., A-Authors:Voorhees_J.J., A-Authors:Nair_R.P., A-Authors:Jenisch_S., A-Authors:Christophers_E., A-Authors:Elder_J.T., A-Authors:Chia_N.V.C., A-Authors:Lim_H.W., A-Authors:Stuart_P.E., A-Authors:Nistor_I., A-Authors:Hiremagalore_R., A-Authors:Weichenthal_M.]->[D-Disease:TopicPSORIASIS 1; SUSCEPTIBILITY TO;

Fig. 3.2 The WAR visualization interface.

is to discover the collaborations among groups of researchers who have conducted the most influential studies on genomics or genetics from OMIM publication data and the related citations/topics.

Data sources. One of the most comprehensive and authoritative compendium of human genes and genetic phenotypes is the Online Mendelian Inheritance in Man (OMIM) database [8]. OMIM is part of the National Center for Biotechnology Information (NCBI) system of databases [7] and it is freely available on the Web. OMIM collects information on all known Mendelian disorders and over 12,000 genes. Specifically, it describes the relationships between phenotypes and genotypes by providing full-text, referenced overviews on genetic disorders. The database is updated daily and thus its content is continuously growing over time. The Application Programming Interfaces (APIs) of OMIM are accessible by public for genetic data crawling and download. These APIs allow users to download the list of all known disorders and a set of related annotations. Disorder annotations consist of (i) a list of scientific publications ranging over the disorder (for each publication the complete bibliographic information is known), (ii) a textual description of the disorder including references, and (iii) links to other genetics resources. To crawl data from the online OMIM database, we considered the exposed APIs [8]. To retrieve the number of citations received by each publication in OMIM we considered the APIs of the PubMed digital library [7]. The integrated dataset contains 8825 articles, 34555 authors, and 302 disorders which were obtained by integrating publication data crawled from OMIM and citation data crawled from PubMed,

Prepared datasets. For each publication in OMIM, the related topics can be extracted in two ways: from metadata (i.e., the descriptions of the genetic disorders associated with the publication) or from the Author-Topic Model (see Section 3.2.2). However, part of the OMIM publications have no full-text access through the exposed APIs. Therefore, we enriched all publications in OMIM with topics extracted from metadata, while we applied the ATM to extract 10 topics only for the subset of the publications in OMIM for which the full-text is available. For the sake of simplicity, from now we will denote as *Disorder* the dataset collecting OMIM publication, disorder topics, and citation counts, while we will denote as *ATM* the dataset collecting the portion of OMIM publication with free full-text version, the related citations, and the ATM main topics. For each paper of *Disorder*, one single disorder topic per paper is available. Differently, for the *ATM* dataset, we selected the top 5 most related topics for each paper, based on the output of the ATM algorithm.

Comparison betwee OMIM disorders and ATM topics. We analyzed the similarity between the ten automatically extracted ATM topics and the manually assigned 302 OMIM disorders in the analyzed publication data. Specifically, we first analyzed the distribution of the OMIM disorders within each subset of publications related to the same topic. Most topics appeared to be almost uncorrelated with the OMIM disorders, as the most frequent disorders typically occurred in no more than 5% of the publications of a given topic. Furthermore, OMIM disorders are associated with 80%-90% of the ATM topics. Hence, the two categorizations seem to be not correlated with each other, as they were generated in different ways and with completely different purposes.

This section is organized as follows. Section 3.3.1 reports some examples of WARs belonging to different categories, which allowed us to answer to the research questions posed in the previous sections. In the subsquent sections, a quantitative analysis of the mining results is reported. Specifically, we discuss (i) the accuracy of the mined rules in identifying the main topics covered by a set of researchers (Section 3.3.2.1), (ii) the distribution of the extracted WARs in the selected categories (Section 3.3.2.2), (iii) the impact of the parameter settings on the number of extracted WARs (Sections 3.3.2.3-3.3.2.4)

3.3.1 Knowledge discovery from the mined WARs

In this section, some examples of WARs are reported separately for each category and we illustrate how these patterns can be exploited to answer to the questions posed in the previous sections (see Section 3.2).

Category (1) comprises Authors-Topic weighted rules (A-T WARs). They can be used to answer to the following questions:

On what topics each collaboration focused on?

Which are the most fruitful authors' collaborations?

Is the authors' collaboration focused on a specific topic or spread over multiple topics?

Table 3.2 reports the top 5 Authors-Topic rules (A-T WARs), in sorting order of decreasing wsup, mined from Disorder. Each A-T rule indicates a specific set of authors who have collaborated on a particular topic. Rule profitability was measured in terms of number of citations received by the co-authored publications. In fact, a high wsup value implies a high number of citations for the papers co-authored by the set of authors reported in the antecedent of the A-T rule. For example, we discover, based on the extracted WARs, that authors Siddique T. and Deng H. X. wrote a set of papers on the Amyotrophic lateral sclerosis disorder and their co-authored publications have been cited 1861 times. Since this WAR is the most frequent one among all the mined A-T WARs ranging over the topic, we can conclude that

Siddique T. and Deng H. X. are among the most influential/authoritative group of researchers about Amyotrophic lateral sclerosis.

A-T rule	wsup	wconf
{(*Author*:Bignell, G.R.), (*Author*:Davies, H.), (*Author*:Garnett, M.J.), (*Author*:Cox, C.), (*Author*:Stephens, P.), (*Author*:Edkins, S.), (*Author*:Clegg, S.), (*Author*:Teague, J.), (*Author*:Woffendin, H.), (*Author*:Bottomley, W.), (*Author*:Davis, N.), (*Author*:Dicks, E.)} → {(*Topic*:MELANOMA CUTANEOUS MALIGNANT SUSCEPTIBILITY TO 1)}	1861	100%
(*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%

Table 3.2 *Disorder dataset*: Top 5 Authors-Topic rules (A-T WARs) in terms of wsup

In Table 3.2, the most frequent A-T WAR is associated with a relatively large group of authors, which consists of 12 different authors. This is typical in the medical domain for which papers are usually co-authored by a large number of authors.

In Table 3.2, all the WARs are characterized by maximal confidence value (100%). This means that the set of authors appearing in the rule antecedent have collaborated only on the topic reported in the consequent of the associated rule. For instance, Siddique T. and Deng H. X. have fruitful collaborations on the Amyotrophic lateral sclerosis disease but they have not produced significant literature on any other topics (according to our data-driven analyses). However, the authors who have had fruitful collaborations on a specific topic are likely to collaborate on other topics as well. To investigate whether authors' collaborations are focused on a specific topic or spread over multiple topics we can compare the A-T WARs characterized by the same antecedent by considering their confidence values as well. For example, Table 3.3 reports four WARs that can be exploited to characterize the collaborations between Brown, E.M. and Kifor, O. and those between Seidman, J.G. and Seidman, C.. Specifically, the first two A-T WARs reported in Table 3.3 show that Brown, E.M. and Kifor, O. have had fruitful collaborations on two main topics: HYPOCALCIURIC HYPERCALCEMIA FAMILIAL TYPE I and HYPOCALCEMIA AUTOSOMAL DOMINANT 1. Their papers on the first topic have received 79.4% of their overall citations (by considering only the co-authored publications), while the second topic is associated with 20.6% of their citations. The sum of the two rule confidence values is 100%. Hence, Brown, E.M. and Kifor, O. have collaborated only on the aforesaid topics. Table 3.3 reports the last two rules that can be used to characterize the collaborations between other two researchers. Based on the mined rules Seidman, J.G. and Seidman, C. have had profitable collaborations on two topics (ARDIOMYOPATHY FAMILIAL HYPERTROPHIC 1 and CARDIOMYOPATHY DILATED 1A). However, since the sum of the confidence values of those rules is less than 100%, we can deduce that Seidman, J.G. and Seidman, C. have co-authored papers on other topics as well, but the latter works have not received a sufficiently high number of citations to be deemed as "relevant" (i.e., no other A-T WARs with wsup above 50 and wconf above 50% associated with Seidman, J.G. and Seidman, C. were mined).

A-T rule	wsup	wconf
{(*Author*:Brown, E.M.), (*Author*:Kifor, O.)} → {(*Topic*:HYPOCALCIURIC HYPERCALCEMIA FAMILIAL TYPE I)}	485	79.4%
{(*Author*:Brown, E.M.), (*Author*:Kifor, O.)} → {(*Topic*:HYPOCALCEMIA AUTOSOMAL DOMINANT 1)}	126	20.6%
{(*Author*:Seidman, J.G.), (*Author*:Seidman, C.)} → {(*Topic*:CARDIOMYOPATHY FAMILIAL HYPERTROPHIC 1)}	566	52.8%
{(*Author*:Seidman, J.G.), (*Author*:Seidman, C.)} → {(*Topic*:CARDIOMYOPATHY DILATED 1A)}	196	18.3%

Table 3.3 *Disorder dataset*: Examples of A-T WARs describing authors who have collaborated on multiple topics

AT-A rule	wsup	wconf
{(*Author*:Rioux, J.D.), (*Author*:Silverberg, M.S.), (*Topic*:INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1)} → {(*Author*:Barmada, M.M.)}	1385	94.2%
{(*Author*:Rioux, J.D.), (*Author*:Silverberg, M.S.), (*Topic*:INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1)} → {(*Author*:Bitton, A.)}	852	57.9%

Table 3.4 *Disorder dataset*: Examples of AT-A WARs

The analysis is extended in order to analyze the collaborations between the aforesaid groups and other researchers. Specifically, we want to answer to the following question:

"*Working on a given topic, has the group (occasionally) collaborated with external authors?*".

The AuthorsTopic-Author rules (AT-A WARs) can support experts in tackling this issue. Table 3.4 reports two example AT-A WARs that can be used to discover who have collaborated with Rioux, J.D. and Silverberg, M.S. on topic INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1 disease (the interest of the group on the specific topic were previously discovered by analyzing the fourth A-T WAR reported in Table3.2). According to the AT-A WAR, Rioux, J.D. and Silverberg, M.S. have conducted joint works with Barmada, M.M. and Bitton, A. on the INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1 topic. Specifically, 94.2% of their citations on that topic are due to papers co-authored by Barmada, M.M. as well, while Bitton, A. has co-authored papers associated with 57.9% of their citations.

A-AT rule	**wsup**	**wconf**
{(*Author*:Almer, S.), (*Author*:Finkel, Y.)} → {(*Author*:Colombel, J.-F.), (*Topic*:INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1)}	67	6.2%
{(*Author*:Cho, J.H.), (*Author*:Brant, S.R.)} → {(*Author*:Bayless, T.M.), (*Topic*:INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1)}	140	15.4%

Table 3.5 *Disorder dataset*: Examples of A-AT WARs

Furthermore, another analysis is performed to analyze the collaborations between a group of researchers and "external" researchers and discover the topics of these collaborations. Specifically we are interested in answering to the question:

"*Has the group (occasionally) collaborated with external authors? On which topics?*"

Table 3.5 reports some examples of Authors-AuthorTopic rules (A-AT WARs). They can be used in order to answer the questions reported above. Based on the mined rules, the group of authors Almer, S. and Finkel, Y. has frequently collaborated only with Colombel, J.-F. on the INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1 topic. Moreover, this collaboration has covered only the 6.2% of their total citations (independently of the topics of the papers co-authored by Almer, S. and Finkel, Y.). Hence, authors Almer, S. and Finkel, Y. seem to have had a limited collaborations with researches external to their group. In Table 3.5, the second rule shows the "external" collaboration of the set of authors Cho, J.H. and Brant, S.R.. Even this group of authors has had an 'external" collaboration with another researcher (Bayless, T.M.) and the target of the collaboration was the INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1 topic.

In order to analyze the AuthorsTopics-Topic rules (AT-T WARs), another analysis is considered to characterize the cross-topic collaborations among authors when pa-

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

Table 3.6 *ATM dataset*: ATM topis

AT-T rule	wsup	wconf
$\{$(*Author*:Shelbourne, P.), (*Author*:Davies, J.), (*Author*:Johnson, K.), (*Topic*:T8)$\} \rightarrow \{$(*Topic*:T9)$\}$	466	100%
$\{$(*Author*:Johnson, K.), (*Author*:Buxton, J.), (*Topic*:T6)$\} \rightarrow \{$(*Topic*:T8)$\}$	456	100%

Table 3.7 *ATM dataset*: Examples of AT-T WARs

pers are characterized by multiple topics. Specifically, we are interested in answering to the question:

"*Given a set of co-authors collaborating on a set of topics, which other topic is likely to be covered by their co-authored publications?*"

As the Disorder dataset contains a single topic per paper, in the following we will consider the AT-T WARs extracted from the ATM dataset as representative example (see Table 3.7). For example, based on the first rule reported in Table 3.7, we can state that 100% of the publications related to topic T8 co-authored by Shelbourne, P., Davies, J., Johnson, K., Shelbourne, P., Davies, J., and Johnson, K. cover also Topic T9. Hence, the publications of the reported co-authors related to topic T8 are also related to topic T9 (i.e., those publications are related to the cross-topic collaboration on topics T8 and T9). Similar considerations hold for the second example AT-T WAR. For the sake of completeness, Table 3.6 reports the top 10 most related terms extracted by the ATM algorithm for each of the identified topics.

Independently of the authors, we could be interested in analyzing the correlations among multiple topics to understand if the same topics are frequently covered by the same publication. Specifically, we are interested in answering to the question:

T-T rule	wsup	wconf
{(*Topic*:T3), (*Topic*:T5), (*Topic*:T6), (*Topic*:T8)} → {(*Topic*:T7)}	1326	95.5%
{(*Topic*:T2), (*Topic*:T5), (*Topic*:T8), (*Topic*:T9)} → {(*Topic*:T4)}	1449	93.4%
{(*Topic*:T2)} → {(*Topic*:T0)}	2118	27.8%
{(*Topic*:T5)} → {(*Topic*:T0)}	2205	23.8%

Table 3.8 *ATM dataset*: Examples of T-T WARs

"*Given a set of publications related to a particular subset of topics, which other topic is also frequently covered in those publications?*"

Table 3.8 shows examples of Topics-Topic rules (T-T WARs), which can be used to identify frequent correlations among topics. Specifically, Table 3.8 reports the top two most confident T-T WARs mined from the ATM dataset and the two less confident ones. The mined WARs show that single topics are usually not very correlated with each other (i.e., the last two rules are both characterized a low confidence value), while the publications covering a large set of topics can be highly correlated with a further topic (see the first two rules reported in Table 3.8). This result is consistent with the main goal of the ATM algorithm, which aims at identifying orthogonal topics (i.e., couples of topics are likely to be weakly correlated). Based on the last two T-T WARs reported in Table 3.8, it turns out that T2 is not very correlated with T0, and T5 is almost uncorrelated with T0 as well. The aforesaid considerations are consistent with the results of a qualitative comparison between the corresponding word-based topic descriptions in Table 3.6.

3.3.2 Quantitative analysis of the characteristics of mined WARs and performance of CSCA

The goal of this section is manifold. Specifically,

(i) We report a quantitative assessment of the reliability of the mined WARs on publication data (see Section 3.3.2.1).

(ii) We analyze the per-length and per-category WAR distributions by setting a standard configuration for the WAR mining algorithm (see Section 3.3.2.2).

(iii) We discuss the impact of the algorithm parameter settings on the quality of the mining results (see Section 3.3.2.3 and 3.3.2.4).

(iv) We discuss the complexity of the CSCA system and we evaluate system performance in terms of execution time (see Section 3.3.2.5).

3.3.2.1 Quantitative assessment of the correctness of the mined WARs

A quantitative assessment of the reliability is performed on the mined WARs. This validation phase focused on A-T WARs and separately, for each dataset,the top 50 WARs were selected by decreasing weighted confidence. The goal in this validation process is to estimate to what extent each of the mined rules is relevant by measuring the pertinence of the topic recommended by the rule head with those of the most influential studies of the authors indicated in the rule body. Specifically, for each A-T WAR *r* we compared the topic in the rule *r*'s consequent with those of the top 3 most cited publications of each author in the rule antecedent. Then, we defined as *score* of rule *r* the percentage of authors who published at least one of his top cited publications on the rule topic. This measure indicates the extent to which the authors mentioned in the rule have the assigned topic in their expertise. A high rule score indicates that the co-occurrence between multiple authors and the topic, which were extracted from publication data based on citation counts, is unlikely to be generated by chance as they reflect the expected single author-topic dependencies.

The average score was 99.8% for the Disorder dataset and 97.5% for the ATM dataset, respectively. This result confirms that the extracted author-topic associations can be deemed as reliable.

3.3.2.2 Characteristics of the mined WARs

In order to analyze the characteristics of the mined WARs, we first set, as standard configuration, the minimum weighted support threshold (i.e., the least citation count value) to 50 and the minimum weighted confidence threshold (i.e., the minimum percentage of publications for which the implication holds) to 50%. The impact of the aforesaid parameters will be discussed later.

Figures 3.3 and 3.4 plot the number of WARs per category (see Section 3.2.4.1) mined from the *Disorder* and *ATM* datasets. As expected, the number of A-T WARs is significantly lower than those of the other ones, because the number of possible combinations is usually at least one order of magnitude lower. The distributions of AT-A and A-TA WARs are approximately the same when only one topic per article

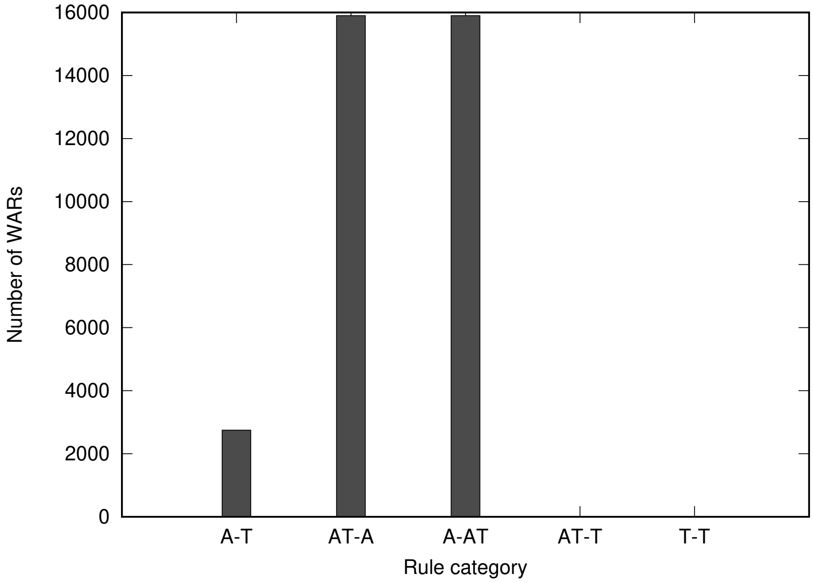


Fig. 3.3 *Disorder dataset*: Distribution of WARs per category. wsup=50, wconf=50%.

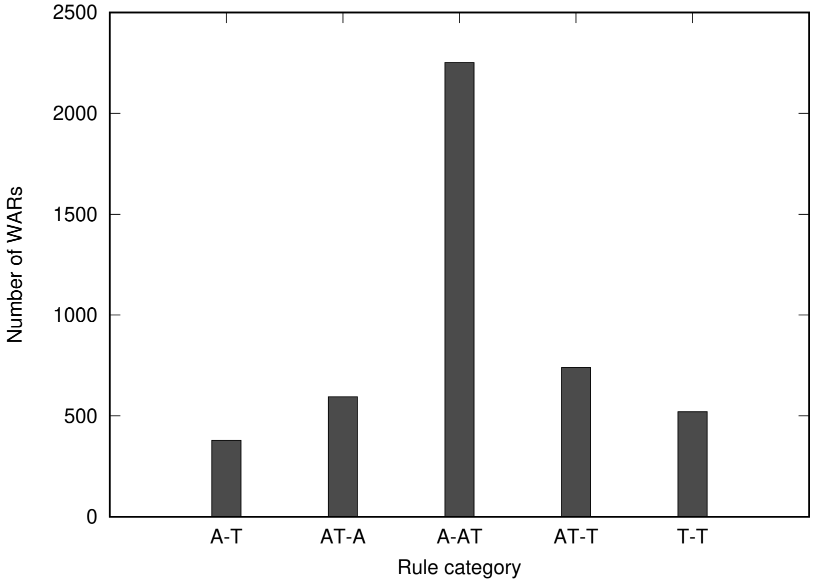


Fig. 3.4 *ATM dataset*: Distribution of WARs per category. wsup=50, wconf=50%.

is available (i.e., for the *Disorder* dataset) because they are generated from the same closed itemset by permuting the corresponding items.

For each category, we analyzed also the per-length distribution of the corresponding WARs (i.e., the number of contained items). As representative examples, Figures 3.5-3.9 report the per-length distribution of WARs of different categories

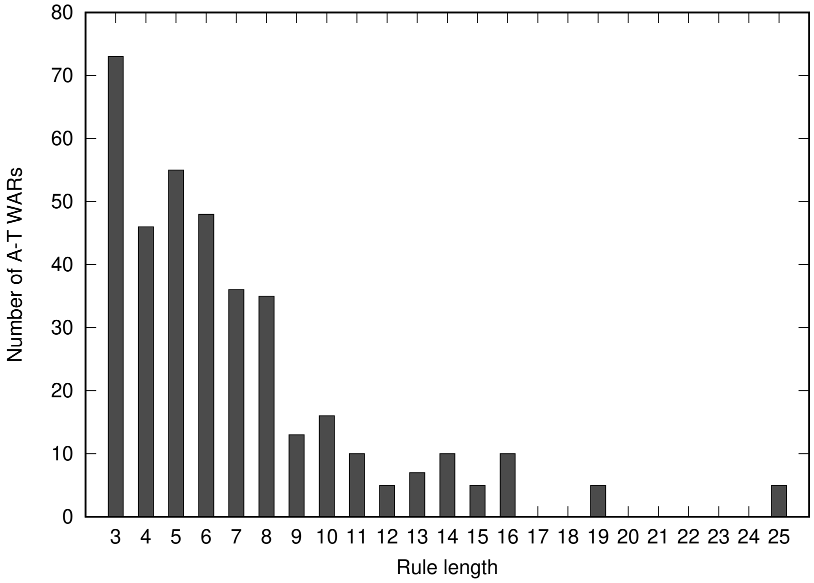


Fig. 3.5 *ATM dataset*: Distribution of A-T WARs per length. wsup=50, wconf=50%.

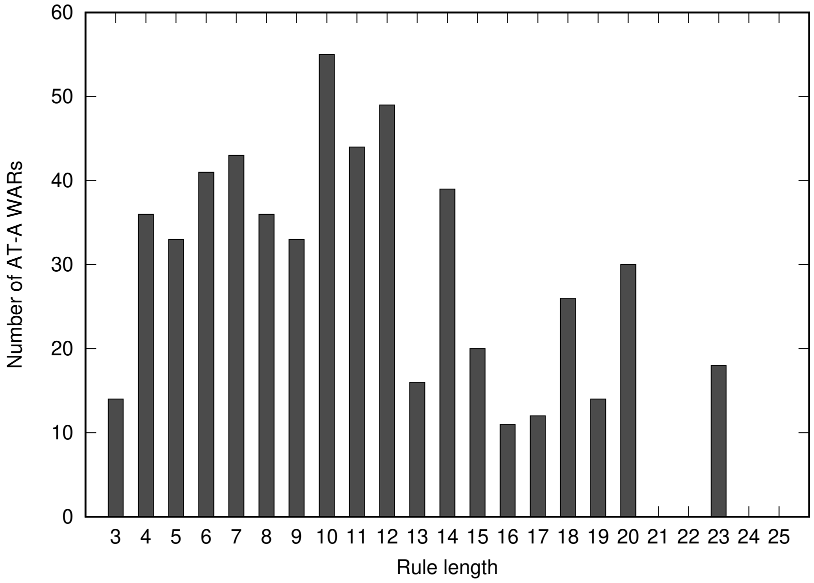


Fig. 3.6 *ATM dataset*: Distribution of AT-A WARs per length. wsup=50, wconf=50%.

mined from *ATM*. We selected the rules mined from the *ATM* dataset because ATM is characterized by multiple topics for each paper and hence WARS of all categories are mined.

Shorter WARs (i.e., WARs with few authors and a topic) within all categories are more numerous than longer ones, because they are most likely to satisfy the support

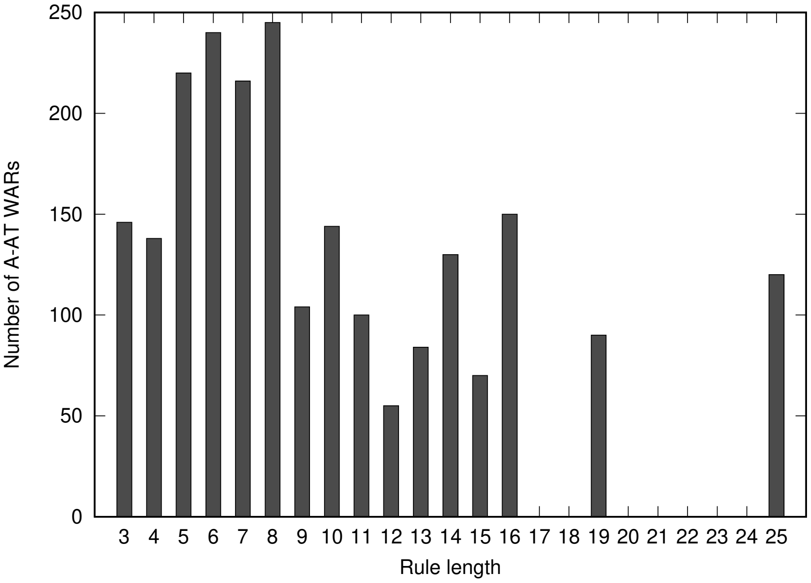


Fig. 3.7 *ATM dataset*: Distribution of A-AT WARs per length. wsup=50, wconf=50%.

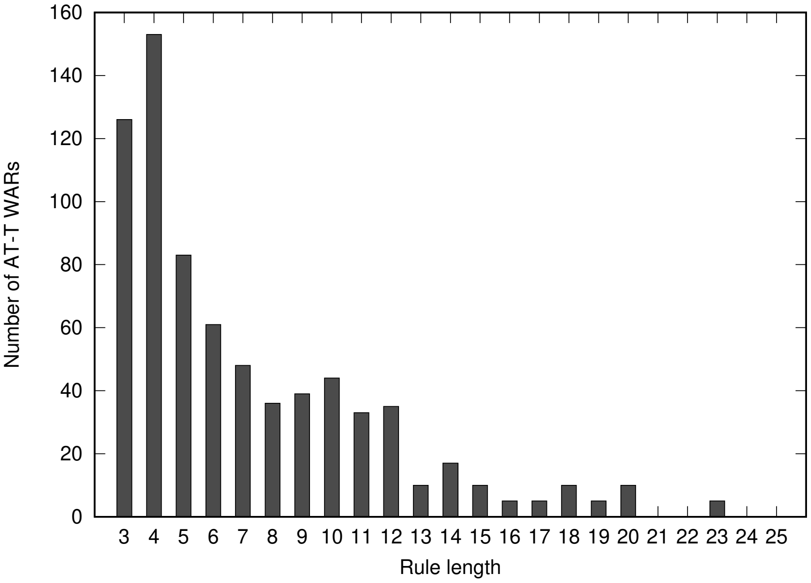


Fig. 3.8 *ATM dataset*: Distribution of AT-T WARs per length. wsup=50, wconf=50%.

threshold. However, as discussed in Section 3.3.1, long WARs provide interesting information about large research groups, which cannot be easily inferred from the Author-Topic Model [53].

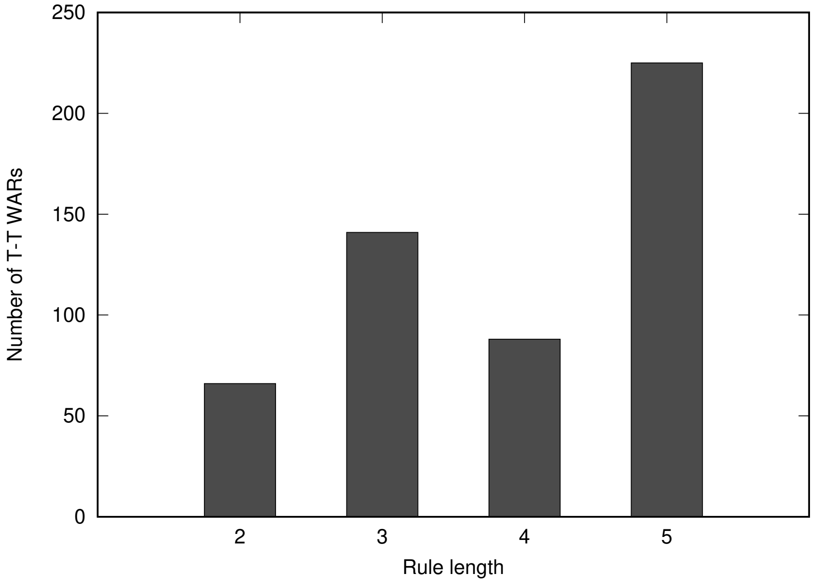


Fig. 3.9 *ATM dataset*: Distribution of T-T WARs per length. wsup=50, wconf=50%.

3.3.2.3 Impact of the minimum weighted support threshold

Figures 3.10 and 3.11 show the cumulative distribution of the number of A-T WARs (chosen as representative) mined from the *Disorder* and *ATM* datasets, respectively, by varying the value of the weighted support threshold. The plots were generated by counting the number of A-T WARs for each distinct value of wsup while keeping the value of wconf fixed to its standard value (50%).

As expected, the number of mined WARs decreases while considering higher wsup values.

3.3.2.4 Impact of the minimum weighted confidence threshold

Figures 3.12 and 3.13 report the cumulative distribution of the number of A-T WARs (chosen as representative) mined from the *Disorder* and *ATM* datasets, respectively, by varying the value of the weighted confidence threshold. The plots were generated by counting the number of A-T WARs for each distinct value of wconf while keeping the value of wsup fixed to its standard value (50).

The results show that the confidence threshold is not very selective, because most of the mined WARs have fairly high confidence (less than 20% of the WARs have wconf below 80%). This is due to the highly influential works on a single topic that

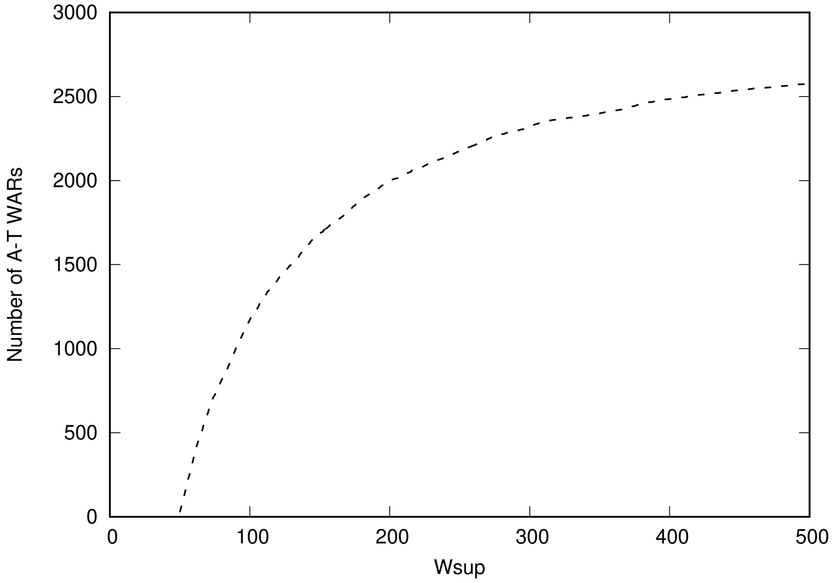


Fig. 3.10 *Disorder dataset*: Cumulative A-T WAR distribution w.r.t. wsup. wconf=50%.

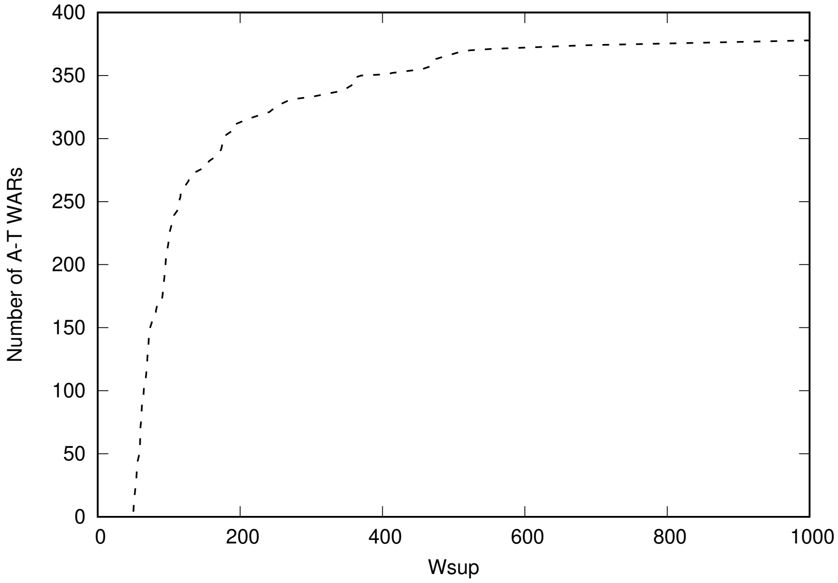


Fig. 3.11 *ATM dataset*: Cumulative A-T WAR distribution w.r.t. wsup. wconf=50%.

most groups of researchers have produced, so the confidence of the corresponding rule is very high. Conversely, the confidence of A-T WARs decreases in case a group has produced scientific works on many different topics. Notice that since publications are weighted by the corresponding number of citations, the collaborations that did not produce any influential works are automatically penalized.

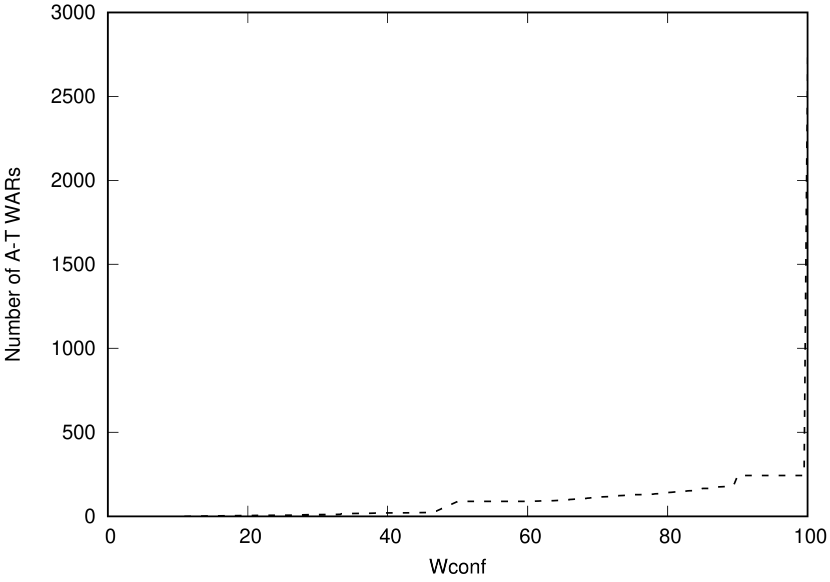


Fig. 3.12 *Disorder dataset*: Cumulative A-T WAR distribution w.r.t. wconf. wsup=50.

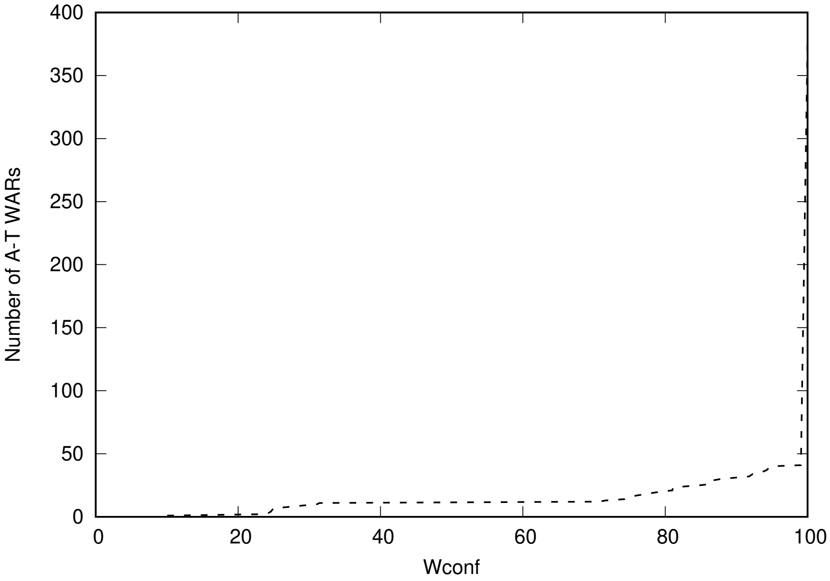


Fig. 3.13 *ATM dataset*: Cumulative A-T WAR distribution w.r.t. wconf. wsup=50.

3.3.2.5 Complexity and execution time

We experimentally analyzed the execution time spent by our approach on *ATM* and *Disorder* datasets.

The most computationally intensive tasks are (i) ATM topic detection and (ii) WAR mining. Data preparation and WAR ranking have negligible impact of the execution time. The time complexity of ATM topic detection is of order of the total number of word tokens in the analyzed dataset multiplied by the number of topics. On the *ATM* dataset each run of the generative process takes approximately 20s. This step is not needed on *Disorder* as topics were directly extracted from publication metadata.

The WAR mining process has linear complexity with respect to the number of mined (closed) itemsets, which, in turn, is combinatorial with the number of items ($2^{\#items}$ in the worse case) [75]. Therefore, the time complexity is super-linear with the number of word tokens in the publication documents. For example, on the *Disorder* dataset the WAR mining process took approximately 35s with wsup=100 (approximately 3700 mined WARs), 238s with wsup=50 (21000 mined WARs), and 998s with wsup=25 (23200 mined WARs).

We compared also the performance of the WAR mining process based on closed itemsets with that of a variant of the original process based on all the frequent itemesets (including non-closed itemsets). By relaxing the constraint on closed itemset mining, more than 100 millions of frequent itemsets were generated from both the Disorder and ATM datasets by enforcing a relatively high wsup value (100). The number of the mined frequent itemsets is at least three orders of magnitude larger than those of closed itemsets. The rule generation process on top of frequent itemsets did not terminate due to the huge number of candidate rule combinations (7GB of itemsets for the Disorder dataset, more than 15 GB for the ATM dataset). Therefore, the WAR mining and exploration process becomes practically unfeasible. The reason is that since many articles have a large number of authors, extracting all the frequent itemsets would generate a huge number of redundant patterns. Conversely, closed itemsets represent a more compact representation of the data recurrences.

Chapter 4

Conclusions

The thesis, first addressed the video hyperlinking problem by proposing enriched query formulations and their combinations. The features considered in the proposed approaches are textual and include ASR transcripts, visual concepts and video metadata, enriched with Named-Entity Recognition and a concept-mapping technique. Experiments addressed the parameter impacts of the different components involved in the query enrichment process and results from the TRECVID submission of the proposed approaches. In particular, the Automatic Feature Selection (AFS) approach reached higher performance than all the other TRECVID competitors for the specific video hyperlinking task.

Detailed analysis of such contributions highlighted that each monomodal query is specifically useful for a subset of the test anchors. Hence, approaches (i) considering ensembles of different monomodal queries and (ii) able to let the best specific query emerge for each test anchor, yielded the best overall results consistently across different metrics.

The second part of this thesis addresses the problem of discovering and ranking fruitful cross-topic collaborations among researchers. The aim is to characterize each research collaboration by discovering the main topics covered and their relative importance in terms of attention given by the research community. To address this issue, a data mining-oriented methodology is proposed, which relies on weighted association rule-based techniques.

The experiments, which were conducted on PubMed and OMIM databases, highlight cross-topic collaborations among multiple authors which cannot be easily inferred using traditional models (e.g., the ATM by Rosen-Zvi et al. [54]).

References

[1] Mohammad Reza Kavoosifar, Daniele Apiletti, Elena Baralis, Paolo Garza, and Benoit Huet. Effective video hyperlinking by means of enriched feature sets and monomodal query combinations. *International Journal of Multimedia Information Retrieval*, pages 1–13, 2019.

[2] Benoit Huet, Elena Baralis, Paolo Garza, and Mohammad Reza Kavoosifar. Eurecom-Polito at TRECVID 2017: Hyperlinking task. In *Working Notes of the TRECVID 2017 Workshop*, 2017.

[3] George Awad, Asad Butt, Jonathan Fiscus, David Joy, Andrew Delgado, Martial Michel, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Quénot, Maria Eskevich, Roeland Ordelman, Gareth J. F. Jones, and Benoit Huet. TRECVID 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking. In *Proceedings of TRECVID 2017*. NIST, USA, 2017.

[4] Jean-Luc Gauvain. The quaero program: Multilingual and multimedia technologies. In *International Workshop on Spoken Language Translation (IWSLT)*, 2010.

[5] Lori Lamel. Multilingual speech processing activities in quaero: Application to multimedia search in unstructured data. In *Baltic HLT*, pages 1–8, 2012.

[6] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[7] NCBI. National Center for Biotechnology Information Website. Available at http://www.ncbi.nlm.nih.gov/ Last access: May 2017, 2017.

[8] Ada Hamosh, Alan F. Scott, Joanna Amberger, David Valle, and Victor A. McKusick. Online mendelian inheritance in man (omim). *Human Mutation*, 15(1):57–61, 2000.

[9] Wei Wang, Jiong Yang, and Philip S. Yu. Efficient mining of weighted association rules (WAR). In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD'00*, pages 270–274, 2000.

[10] Guidelines for TRECVID 2016, Video Hyperlinking (LNK) task, available at https://www-nlpir.nist.gov/projects/tv2016/tv2016.html.

[11] Hierarchical decomposition and representation of video content, available at http://www.scholarpedia.org/article/video_content_structuring.

[12] Maria Eskevich, Gareth JF Jones, Robin Aly, Roeland JF Ordelman, Shu Chen, Danish Nadeem, Camille Guinaudeau, Guillaume Gravier, Pascale Sébillot, Tom De Nies, et al. Multimedia information seeking through search and hyperlinking. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 287–294. ACM, 2013.

[13] Maria Eskevich, Robin Aly, David Racca, Roeland Ordelman, Shu Chen, and Gareth JF Jones. The search and hyperlinking task at mediaeval 2014. 2014.

[14] Petra Galuščáková, Shadi Saleh, and Pavel Pecina. Shamus: Ufal search and hyperlinking multimedia system. In *European Conference on Information Retrieval*, pages 853–856. Springer, 2016.

[15] Mohammad Soleymani, Michael Riegler, and Pål Halvorsen. Multimodal analysis of user behavior and browsed content under different image search intents. *International Journal of Multimedia Information Retrieval*, 7(1):29–41, 2018.

[16] Akihiko Nakagawa, Andrea Kutics, Kiyotaka Tanaka, and Masaomi Nakajima. Combining words and object-based visual features in image retrieval. In *Image Analysis and Processing, 2003. Proceedings. 12th International Conference on*, pages 354–359. IEEE, 2003.

[17] Ariadna Quattoni, Michael Collins, and Trevor Darrell. Learning visual representations using images with captions. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[18] Xueliang Liu, Raphaël Troncy, and Benoit Huet. Finding media illustrating events. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 58. ACM, 2011.

[19] Tomoki Okuoka, Tomokazu Takahashi, Daisuke Deguchi, Ichiro Ide, and Hiroshi Murase. Labeling news topic threads with wikipedia entries. In *Multimedia, 2009. ISM'09. 11th IEEE International Symposium on*, pages 501–504. IEEE, 2009.

[20] Anthony Hoogs, AG Amitha Perera, Roderic Collins, Arslan Basharat, Keith Fieldhouse, Chuck Atkins, Linus Sherrill, Benjamin Boeckel, Russell Blue, Matthew Woehlke, et al. An end-to-end system for content-based video retrieval using behavior, actions, and appearance with interactive query refinement. In *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*, pages 1–6. IEEE, 2015.

[21] Adam Blažek, Jakub Lokoč, Filip Matzner, and Tomáš Skopal. Enhanced signature-based video browser. In *International Conference on Multimedia Modeling*, pages 243–248. Springer, 2015.

[22] Kai Uwe Barthel, Nico Hezel, and Radek Mackowiak. Navigating a graph of scenes for exploring large video collections. In *International Conference on Multimedia Modeling*, pages 418–423. Springer, 2016.

[23] Claudiu Tanase, Ivan Giangreco, Luca Rossetto, Heiko Schuldt, Omar Seddati, Stephane Dupont, Ozan Can Altiok, and Metin Sezgin. Semantic sketch-based video retrieval with autocompletion. In *Companion Publication of the 21st International Conference on Intelligent User Interfaces*, pages 97–101. ACM, 2016.

[24] Christian Beecks, Klaus Schoeffmann, Mathias Lux, Merih Seran Uysal, and Thomas Seidl. Endoscopic video retrieval: A signature-based approach for linking endoscopic images with video segments. In *Multimedia (ISM), 2015 IEEE International Symposium on*, pages 33–38. IEEE, 2015.

[25] Mohammed Yassine Kazi Tani, Abdelghani Ghomari, Adel Lablack, and Ioan Marius Bilasco. Ovis: ontology video surveillance indexing and retrieval system. *International Journal of Multimedia Information Retrieval*, 6(4):295–316, 2017.

[26] Luca Rossetto, Ivan Giangreco, Claudiu Tănase, and Heiko Schuldt. Multimodal video retrieval with the 2017 imotion system. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 457–460. ACM, 2017.

[27] Anastasia Moumtzidou, Theodoros Mironidis, Evlampios Apostolidis, Foteini Markatopoulou, Anastasia Ioannidou, Ilias Gialampoukidis, Konstantinos Avgerinakis, Stefanos Vrochidis, Vasileios Mezaris, Ioannis Kompatsiaris, et al. Verge: a multimodal interactive search engine for video browsing and retrieval. In *International Conference on Multimedia Modeling*, pages 394–399. Springer, 2016.

[28] Song Tan, Chong-Wah Ngo, Hung-Khoon Tan, and Lei Pang. Cross media hyperlinking for search topic browsing. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 243–252. ACM, 2011.

[29] Mikail Demirdelen, Mateusz Budnik, Gabriel Sargent, Rémi Bois, and Guillaume Gravier. IRISA at TRECVID 2017: Beyond crossmodal and multimodal models for video hyperlinking. In *Working Notes of the TRECVID 2017 Workshop*, 2017.

[30] Rémi Bois, Vedran Vukotić, Anca-Roxana Simon, Ronan Sicre, Christian Raymond, Pascale Sébillot, and Guillaume Gravier. Exploiting multimodality in video hyperlinking to improve target diversity. In *International Conference on Multimedia Modeling*, pages 185–197. Springer, 2017.

[31] Vedran Vukotić, Christian Raymond, and Guillaume Gravier. A crossmodal approach to multimodal fusion in video hyperlinking. *IEEE MultiMedia*, 25(2):11–23, 2018.

[32] Vedran Vukotić, Christian Raymond, and Guillaume Gravier. Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and crossmodal applications. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 343–346. ACM, 2016.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] Yi Tay, Minh C Phan, Luu Anh Tuan, and Siu Cheung Hui. Learning to rank question answer pairs with holographic dual lstm architecture. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 695–704. ACM, 2017.

[35] Zhi-Qi Cheng, Hao Zhang, Xiao Wu, and Chong-Wah Ngo. On the selection of anchors and targets for video hyperlinking. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 287–293. ACM, 2017.

[36] David Nicolas Racca and Gareth JF Jones. Evaluating search and hyperlinking: An example of the design, test, refine cycle for metric development. In *MediaEval*, 2015.

[37] Yashaswi Verma, Abhishek Jha, and CV Jawahar. Cross-specificity: modelling data semantics for cross-modal matching and retrieval. *International Journal of Multimedia Information Retrieval*, 7(2):139–146, 2018.

[38] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[39] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[40] Sebastian Schmiedeke, Peng Xu, Isabelle Ferrané, Maria Eskevich, Christoph Kofler, Martha A Larson, Yannick Estève, Lori Lamel, Gareth JF Jones, and Thomas Sikora. Blip10000: A social video dataset containing spug content for tagging and retrieval. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 96–101. ACM, 2013.

[41] Benoit Huet, Elena Baralis, Paolo Garza, and Mohammad Reza Kavoosifar. Eurecom-Polito at TRECVID 2016: Hyperlinking task. In *Working Notes of the TRECVID 2016 Workshop*, 2016.

[42] Roger B Bradford and John Pozniak. A systematic approach to design of a text categorizer. In *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on*, pages 509–514. IEEE, 2016.

[43] Maria Eskevich, Quoc-Minh Bui, Hoang-An Le, and Benoit Huet. Exploring video hyperlinking in broadcast media. In *Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia*, pages 35–38. ACM, 2015.

[44] Jayant Kumar. *Apache Solr Search Patterns*. Packt Publishing Ltd, 2015.

[45] Robin Aly, Maria Eskevich, Roeland Ordelman, and Gareth JF Jones. Adapting binary information retrieval evaluation metrics for segment-based retrieval tasks. *arXiv preprint arXiv:1312.1913*, 2013.

[46] David Smiley, Eric Pugh, Kranti Parisa, and Matt Mitchell. *Apache Solr enterprise search server*. Packt Publishing Ltd, 2015.

[47] Michael McCandless, Erik Hatcher, and Otis Gospodnetic. *Lucene in action: covers Apache Lucene 3.0*. Manning Publications Co., 2010.

[48] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[49] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[50] Dekang Lin et al. An information-theoretic definition of similarity. In *Icml*, volume 98, pages 296–304. Citeseer, 1998.

[51] Satanjeev Banerjee and Ted Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 136–145. Springer, 2002.

[52] Bahjat Safadi, Mathilde Sahuguet, and Benoit Huet. When textual and visual information join forces for multimedia retrieval. In *Proceedings of International Conference on Multimedia Retrieval*, page 265. ACM, 2014.

[53] Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. *CoRR*, abs/1207.4169, 2012.

[54] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’04, pages 306–315, New York, NY, USA, 2004. ACM.

[55] Peter Mutschke. *Mining Networks and Central Entities in Digital Libraries. A Graph Theoretic Approach Applied to Co-author Networks*, pages 155–166. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[56] M. E. J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Rev. E*, 64:2001, 2001.

[57] Jie Tang, Jing Zhang, Limin Yao, Juan-Zi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In *KDD*, 2008.

[58] Scott White and Padhraic Smyth. Algorithms for estimating relative importance in networks. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 266–275, New York, NY, USA, 2003. ACM.

[59] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Seventh International World-Wide Web Conference (WWW 1998)*, 1998.

[60] Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and ChengXiang Zhai. Content-based citation analysis: The next generation of citation analysis. *JASIST*, 65:1820–1833, 2014.

[61] Ha Jin Kim, Juyoung An, Yoo Kyung Jeong, and Min Song. Exploring the leading authors and journals in major topics by citation sentences and topic modeling. In *BIRNDL@JCDL*, 2016.

[62] Guo Zhang, Ying Ding, and Stasa Milojevic. Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. *JASIST*, 64:1490–1503, 2013.

[63] Ngai Meng Kou, Leong Hou U., Nikos Mamoulis, and Zhiguo Gong. Weighted coverage based reviewer assignment. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 2031–2046, New York, NY, USA, 2015. ACM.

[64] Ngai Meng Kou, Leong Hou U, Nikos Mamoulis, Yuhong Li, Ye Li, and Zhiguo Gong. A topic-based reviewer assignment system. *Proc. VLDB Endow.*, 8(12):1852–1855, August 2015.

[65] B. Li and Y. T. Hou. The new automated ieee infocom review assignment system. *IEEE Network*, 30(5):18–24, September 2016.

[66] Luca Cagliero and Paolo Garza. Infrequent weighted itemset mining using frequent pattern growth. *IEEE Trans. Knowl. Data Eng.*, 26(4):903–915, 2014.

[67] Ke Sun and Fengshan Bai. Mining weighted association rules without preassigned weights. *IEEE Transactions on Knowledge and Data Engineering*, 20(4):489 –495, 2008.

[68] Feng Tao, Fionn Murtagh, and Mohsen Farid. Weighted association rule mining using weighted support and significance framework. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD'03*, pages 661–666, 2003.

[69] R. Agrawal, T. Imielinski, and Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD 1993*, pages 207–216, 1993.

[70] Luca Cagliero, Paolo Garza, Mohammad Reza Kavoosifar, and Elena Baralis. Identifying collaborations among researchers: a pattern-based approach. In *Proceedings of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017) co-located with the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017), Tokyo, Japan, August 11, 2017.*, pages 56–68, 2017.

[71] Ch'ao Lu, Chengzhi Zhang, and Shutian Ma. How does citing behavior for a scientific article change over time?: A preliminary study. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, ASIST '15, pages 97:1–97:4, Silver Springs, MD, USA, 2015. American Society for Information Science.

[72] Edward Loper and Steven Bird. NLTK: the Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[73] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[74] Jianyong Wang, Jiawei Han, and Jian Pei. Closet+: searching for the best strategies for mining frequent closed itemsets. In Lise Getoor, Ted E. Senator, Pedro Domingos, and Christos Faloutsos, editors, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 236–245, 2003.

[75] Jiawei Han, Jain Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. *In SIGMOD'00, Dallas, TX*, May 2000.

[76] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th VLDB conference*, pages 487–499, 1994.