xLiMe
FP7-ICT-2013-10
ICT-2013.4.1
Contract no.: 611346
www.xlime.eu

Deliverable D4.1

Statistical Content Linking Prototype

Editor:	Aditya Mogadala, KIT
Author(s):	Aditya Mogadala, KIT
Deliverable Nature:	Prototype (P)
Dissemination Level: (Confidentiality)	Public (PU)
Contractual Delivery Date:	M12 - 31 October 2014
Actual Delivery Date:	M12 - 31 October 2014
Suggested Readers:	Researchers and developers who are interested in doing research and development on the methods used for content comparison across modalities.
Version:	1.0
Keywords:	Cross-modal, cross-lingual, Correlation analysis, kernels, Information retrieval

Disclaimer

This document contains material, which is the copyright of certain xLiMe consortium parties, and may not be reproduced or copied without permission.

In case of Public (PU):
All xLiMe consortium parties have agreed to full publication of this document.

In case of Restricted to Programme (PP):
All xLiMe consortium parties have agreed to make this document available on request to other framework programme participants.

In case of Restricted to Group (RE):
The information contained in this document is the proprietary confidential information of the xLiMe consortium and may not be disclosed except in accordance with the consortium agreement. However, all xLiMe consortium parties have agreed to make this document available to <group> / <purpose>.

In case of Consortium confidential (CO):
The information contained in this document is the proprietary confidential information of the xLiMe consortium and may not be disclosed except in accordance with the consortium agreement.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the xLiMe consortium as a whole, nor a certain party of the xLiMe consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

Full Project Title:	Cross-lingual Cross-Media Knowledge Extraction
Short Project Title:	xLiMe
Number and Title of Work Package:	WP4 Cross-Media Semantic Integration
Document Title:	D4.1 – Statistical Content Linking Prototype
Editor:	Aditya Mogadala, KIT
Work Package Leader:	Aditya Mogadala, KIT

Copyright notice

Executive Summary

The goal of this deliverable is to provide a detailed description of the research and development accomplished to build the prototype for bridging different modalities of data generated from various sources. The prototype provides recommendations of social media links and news articles for the live TV shows telecasted in various channels of internet TV streaming service provided by ZATTOO. Also, it provides recommendations about similar TV shows.

We approach the problem of bridging modalities using two different techniques. The first approach uses only text generated from diverse modalities, while the second one uses heterogeneous features generated from various modalities.

Table of Contents

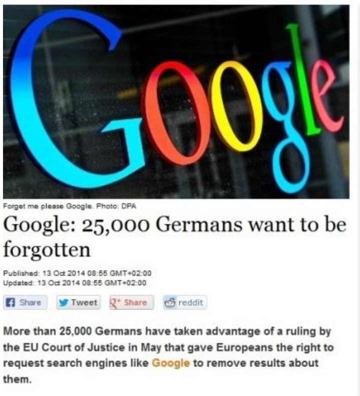
Abbreviations

IPTV	Internet protocol Television
OCR	Optical Character Recognition
RDF	Resource Description Framework
URL	Unified Resource Locator
PTM	Poly-lingual Topic Models
LDA	Latent Dirichlet Allocation
SIFT	Scale invariant Feature Transformation
JSON	Java Script Object Notation

1 Introduction

Most of the data present on the web are distributed into diverse modalities. A modality can be considered as a source of information representing text, image, video or an audio. We consider few examples to understand different modalities present on the web.

First, consider the news articles published online. They generally include one or more modalities. Each article contains either a video or an image along with the text. Similar observations can be made for personal web pages and Wikipedia articles. Figure 1 shows an example article taken from the news publishing site thelocal.de [1] constituting an image and text modality.



Forget me please Google. Photo: DPA

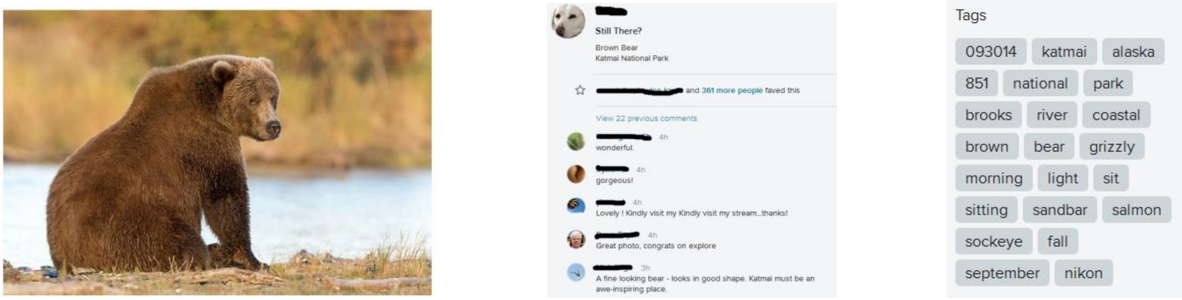
Google: 25,000 Germans want to be forgotten

Published: 13 Oct 2014 08:55 GMT+02:00
Updated: 13 Oct 2014 08:55 GMT+02:00

More than 25,000 Germans have taken advantage of a ruling by the EU Court of Justice in May that gave Europeans the right to request search engines like Google to remove results about them.

Figure 1: News Article representing Image and Text modality

The second example is online photo sharing websites. An uploaded image is usually accompanied by textual description, tags and comments. Figure 2 shows a sample image uploaded to Flickr [2] along with its text represented in various forms. So a photo sharing site usually contains image and text modalities together.



[redacted]

Still There?

Brown Bear
Katmai National Park

☆ [redacted] and 361 more people faved this

View 22 previous comments

[redacted] 4h
wonderful.

[redacted] 4h
gorgeous!

[redacted] 4h
Lovely ! Kindly visit my Kindly visit my stream...thanks!

[redacted] 4h
Great photo, congrats on explore

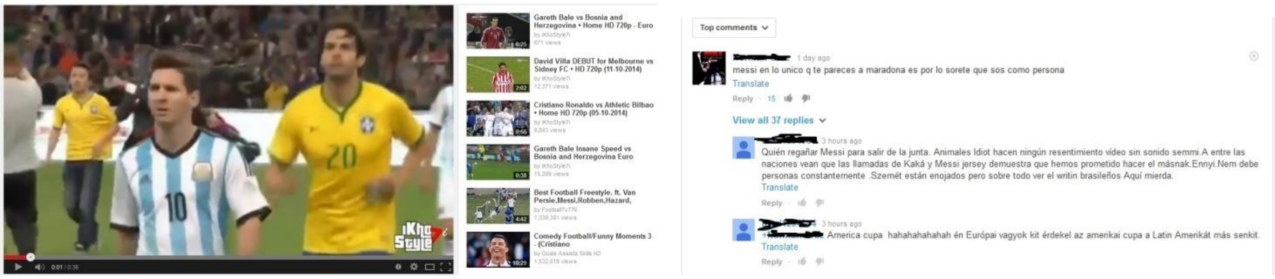
[redacted] 3h
A fine looking bear - looks in good shape. Katmai must be an awe-inspiring place.

Tags

093014 katmai alaska
851 national park
brooks river coastal
brown bear grizzly
morning light sit
sitting sandbar salmon
sockeye fall
september nikon

Figure 2: Image of the brown bear posted on Flickr along with description, tags and comments

The third example is online video publishing websites. Most of the online videos present on publishing websites like YouTube, Vimeo and IPTV services are usually accompanied by video recommendations and textual comments. Figure 3 shows a sample video taken from YouTube [3] along with its video recommendations and textual comments.



Gareth Bale vs Bosnia and Herzegovina • Home HD 720p - Euro
by iKhoStyle7i
671 views

David Villa DEBUT for Melbourne vs Sidney FC • HD 720p (11-10-2014)
by iKhoStyle7i
12,371 views

Cristiano Ronaldo vs Athletic Bilbao • Home HD 720p (05-10-2014)
by iKhoStyle7i
8,843 views

Gareth Bale Insane Speed vs Bosnia and Herzegovina Euro
by iKhoStyle7i
15,289 views

Best Football Freestyle. ft. Van Persie,Messi,Robben,Hazard,
by FootballTv776
1,338,381 views

Comedy Football/Funny Moments 3 - (Cristiano
by Goals Assists Skills HD
1,532,819 views

Top comments

[redacted] 1 day ago
messi en lo unico q te pareces a maradona es por lo sorete que sos como persona
Translate
Reply · 15

View all 37 replies

[redacted] 3 hours ago
Quién regañar Messi para salir de la junta. Animales Idiot hacen ningún resentimiento video sin sonido semmi.A entre las naciones vean que las llamadas de Kaká y Messi jersey demuestra que hemos prometido hacer el másnak.Ennyi.Nem debe personas constantemente .Szemét están enojados pero sobre todo ver el writin brasileños Aquí mierda.
Translate
Reply ·

[redacted] 3 hours ago
[redacted] America cupa hahahahahahah én Európai vagyok kit érdekel az amerikai cupa a Latin Amerikát más senkit.
Translate
Reply ·

Figure 3: YouTube video along with video recommendations and user comments

Multi-modal content shown in the above examples can be used in various applications. Some of them can be used for supporting multimedia search, cross-modal recommendations and computational advertising.

In this deliverable, we address the problem of analyzing multi-modal data generated from various sources. We use the video modality stream generated by ZATTOO [4] IPTV service for providing real-time recommendations to the TV shows. The recommendations are represented as a list of similar text articles published on social media sites and news articles.

The problem of matching different modalities can be seen from the perspective of statistical content matching. Content present in the various modalities can be represented in many ways. Either each modality can be transformed into a common representation of text or can be represented in its native form using raw features.

In our research, we explore both the approaches of using text and raw features for providing content linking. Figure 4 shows the broad Visualization of the approach.

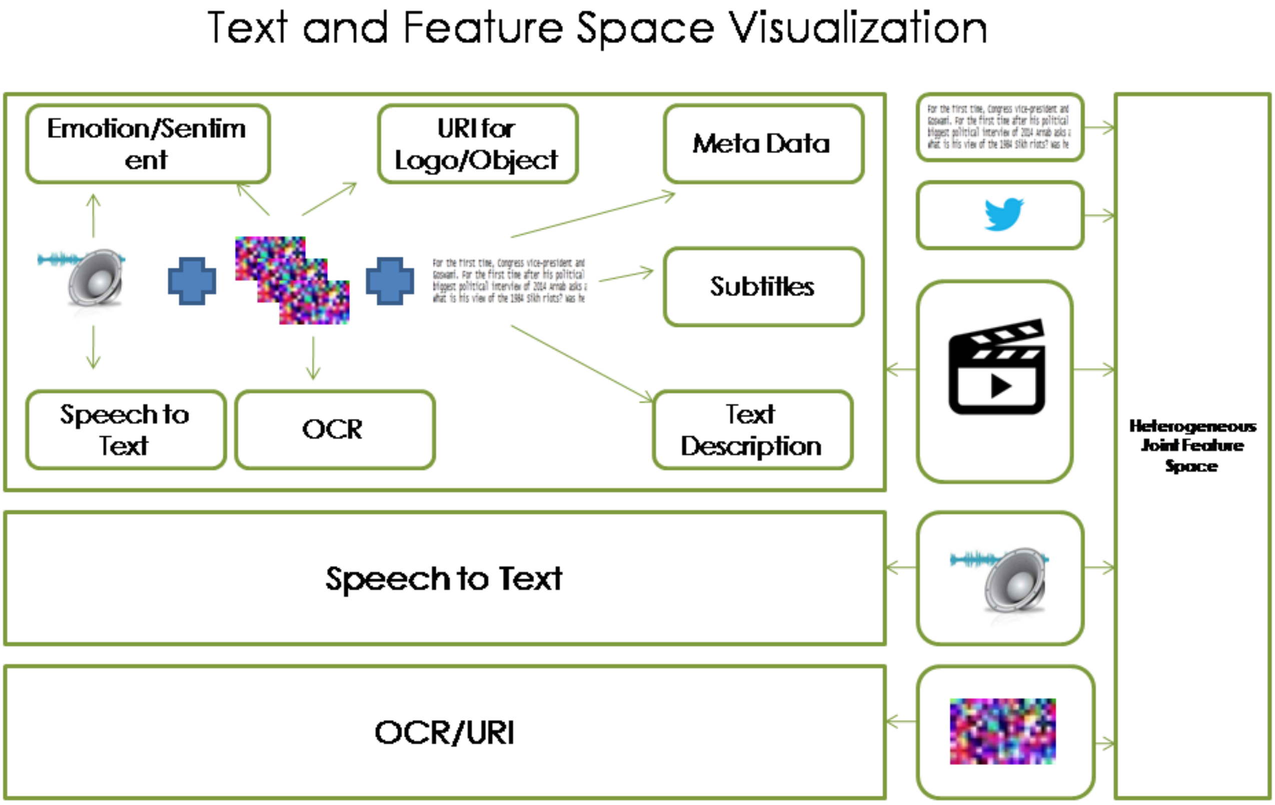


Figure 4: Architecture

It is observed from Figure 4 that image, video, audio and text modalities are either represented by text or by heterogeneous features. Below, we evaluate varied ways in which text can be extracted from these diverse modalities.

- **Image** – If an image contains text, it is extracted using an OCR. Automatic image annotation can be used for assigning metadata in the form of captions or keywords. Further, these keywords can be linked to a knowledge base for facilitating semantic search.
- **Audio** – Speech recognition can be employed to transcribe spoken words into text.
- **Video** – Videos are represented in terms of image frames and the corresponding audio. Sometimes videos may contain text in the form of subtitles to support accessibility. Once the videos are

represented in image and audio modalities, we extract text using the approaches used for image and audio modalities.

- **Social Media Text** – Most of the time text generated in small snippets like Tweets etc., is often noisy. It contains various slangs and spelling errors. Even though the data are represented in the textual modality, it requires certain pre-processing steps for further analysis.

In the following sections, first we present the data generated from different modalities in the xLiMe pipeline and then discuss our approaches briefly.

1.1 Data

In this section, we present the content generated by our use case partners in the format of xLiMe Meta data model [10] from different modalities.

ZATTOO is one of the use case partners who provide IPTV video streams. OCR is performed on the image frames of the ZATTOO video stream to extract text. Below, we can see an example of OCR output encoded in the xLiMe Meta data model.

Example – ZATTOO Video Stream OCR Ouput – xLiMe Meta Data Model

```
{
<http://unitn.it/ocr/0f28b734-c77d-4e95-be6c-9bbee6d0e761>
        dcterms:title        "UNITN OCR" ;
        prov:wasAttributedTo  <http://www.unitn.it> ;
        prov:wasGeneratedBy   [ a                prov:Activity ;
                          prov:endedAtTime    "2014-09-15 14:46:51.369956"^^xsd:date ;
                          prov:startedAtTime  "2014-09-15 14:46:55.369956"^^xsd:date
                        ] .
}

# start of a new UNITN named graph / meta information above
<http://unitn.it/ocr/0f28b734-c77d-4e95-be6c-9bbee6d0e761>
{
   <http://zattoo.com/program/15029881> ma:hasTrack
<http://zattoo.com/program/15029881/video> .
   <http://zattoo.com/program/15029881/video> a ma:VideoTrack ;
   sioc:content """
84260848.000000, Suuiruthcn., GOOGLE, I.-, GOOGLE, DISCUSSES, DISCUSSES, In, 3&1
FEM , , FORGO1, KIUHI, 05:46, KIUI', wllc , 3&1
""" .
}
```

Similarly, audio modality of the ZATTOO video streams is used to generate text from speech. Below, we can see an example of the text extracted from the audio of ZATTOO video streams encoded into xLiMe data model.

Example – ZATTOO Video Stream – (Audio to text) - xLiMe Meta Data Model

```
{
   <http://zattoo.com/processed/15079855> prov:wasGeneratedBy [ a prov:Activity ;
```

```
            dcterms:title "Zattoo ASR" ;
            prov:endedAtTime "2014-09-13T01:36:08+00:00"^^xsd:datetime ;
            prov:startedAtTime "2014-09-13T03:36:04.835207+02:00"^^xsd:datetime ;
            prov:wasAttributedTo <http://ijs.si> ] .
}

<http://zattoo.com/processed/15079855> {
   <http://zattoo.com/program/15079855/audio> a ma:AudioTrack ;
      ma:hasLanguage [ rdfs:label "en" ] ;
      xlime:hasAnnotation [ rdfs:label "BMW" ;
            xlime:hasEntity <http://rdf.freebase.com/ns/m.017yh> ],
         [ rdfs:label "CNBC" ;
            xlime:hasEntity <http://rdf.freebase.com/ns/m.01gl9g> ] ;
      xlime:hasRecognisedSpeech [ xlime:hasASREngine <http://mediaspeech.com/webservice> ;
            xlime:hasASRText " not earnings and revenue relative solid follow his first post fact stares at the
end of the BMW Championship survivor legislators around the city itself which went into my hands and no
one time Yuvraj joining me on CNBC needs website gained exclusive access to some of the world's most
fascinating people what a series of loneliness welcome back to top of the big corp " ;
            xlime:hasStreamPosition 8.420747e+07 ] .
}
```

Now to provide cross-modal recommendations based on content similarity, we need data generated from forums like social media posts and news articles. VICO [5] is another use case partner, who provides social media data generated from facebook posts, tweets, forums and blogs. Below we can see two examples of social media data generated by VICO.

Example – VICO Social Media Data – Blog – xLiMe Meta Data Model

```
{ <http://vico-research.com/social/bd5eff57-8419-44f2-8d84-adae742a8abc>
        dcterms:title        "KIT Annotations" ;
        prov:wasAttributedTo  <http://aifb.kit.edu> ;
        prov:wasGeneratedBy   [ a                prov:Activity ;
                          prov:endedAtTime    "2014-09-01T14:36:55"^^xsd:dateTime ;
                          prov:startedAtTime  "2014-09-01T14:36:36"^^xsd:dateTime
                         ] .
}
<http://vico-research.com/social/http://www.landtreff.de/index.php/2cbb23a7-b6ff-380f-8803-
efa9de52eef7>
         a                 sioc:MicroPost ;
         dcterms:created      "2014-08-25T21:51:00"^^xsd:dateTime ;
         dcterms:language     "de" ;
         dcterms:publisher    <http://www.landtreff.de/index.php> ;
         dcterms:source      <http://www.landtreff.de/post1197936.html#p1197936#5> ;
         dcterms:spatial     [ rdfs:label  "de" ] ;
         sioc:content        "Adidas und Puma" ;
         sioc:has_creator     <http://www.landtreff.de/index.php#WaldbauerSchosi> ;
         xlime:hasAnnotation  [ xlime:hasConfidence  "0.86"^^xsd:double ;
                          xlime:hasEntity      <http://de.dbpedia.org/resource/Puma_%28Unternehmen%29> ;
                          xlime:hasPosition    [ xlime:hasStartPosition  "11"^^xsd:long ;
                                          xlime:hasStopPosition   "15"^^xsd:long
                                         ]
                        ] ;
         xlime:hasAnnotation  [ xlime:hasConfidence  "1"^^xsd:double ;
```

```
                    xlime:hasEntity     <http://de.dbpedia.org/resource/Adidas> ;
                    xlime:hasPosition    [ xlime:hasStartPosition  "0"^^xsd:long ;
                                           xlime:hasStopPosition   "6"^^xsd:long
                                         ]
                  ] .
```

Example – VICO Social Media Data – Tweets –xLiMe Data Model

```
{ <http://vico-research.com/social/f99be0a3-442f-4709-8632-d69bd0509da9>
      dcterms:title        "KIT Annotations" ;
      prov:wasAttributedTo  <http://aifb.kit.edu> ;
      prov:wasGeneratedBy   [ a                prov:Activity ;
                    prov:endedAtTime    "2014-10-15T02:35:50"^^xsd:dateTime ;
                    prov:startedAtTime  "2014-10-15T02:35:45"^^xsd:dateTime
                  ] .
}

<http://vico-research.com/social/f99be0a3-442f-4709-8632-d69bd0509da9> {
  <http://vico-research.com/social/Twitter/7f531a2d-ef60-30b0-a1ca-71f54bb12e34>
       a                  sioc:MicroPost ;
       dcterms:created      "2014-10-08T04:16:06"^^xsd:dateTime ;
       dcterms:language     "en" ;
       dcterms:publisher    <http://www.twitter.com/> ;
       dcterms:source       <http://twitter.com/Daljodh_Singh/statuses/519672553267146752> ;
       dcterms:spatial      [ rdfs:label  "" ] ;
       sioc:content         "RT @_BestRapz: Spilled my drink while playing FIFA, call that a Messi." ;
       sioc:has_creator     <http://twitter.com/Daljodh_Singh> ;
       xlime:hasAnnotation  [ xlime:hasConfidence  "1"^^xsd:double ;
                    xlime:hasEntity     dbpedia:Lionel_Messi ;
                    xlime:hasPosition    [ xlime:hasStartPosition  "64"^^xsd:long ;
                                           xlime:hasStopPosition   "69"^^xsd:long
                                         ]
                  ] ;
       xlime:hasAnnotation  [ xlime:hasConfidence  "0.742"^^xsd:double ;
                    xlime:hasEntity     dbpedia:Drink ;
                    xlime:hasPosition    [ xlime:hasStartPosition  "26"^^xsd:long ;
                                           xlime:hasStopPosition   "31"^^xsd:long
                                         ]
                  ] ;
       xlime:hasAnnotation  [ xlime:hasConfidence  "0.985"^^xsd:double ;
                    xlime:hasEntity     dbpedia:FIFA ;
                    xlime:hasPosition    [ xlime:hasStartPosition  "46"^^xsd:long ;
                                           xlime:hasStopPosition   "50"^^xsd:long
                                         ]
                  ] ;
       xlime:hasAnnotation  [ xlime:hasConfidence  "0.78"^^xsd:double ;
                    xlime:hasEntity     <http://dbpedia.org/resource/RT_%28TV_network%29> ;
                    xlime:hasPosition    [ xlime:hasStartPosition  "0"^^xsd:long ;
                                           xlime:hasStopPosition   "2"^^xsd:long
                                         ]
                  ] .
}
```

News articles are also used for cross-modal recommendations. JSI is another collaborator who collects data from various online news sources and provides content for the xLiMe Meta data model. Below, we can see an example news article provided by JSI.

Example – JSI News Articles – xLiMe Meta Data Model

```
{
  <http://ijs.si/enryched/211017736> prov:wasGeneratedBy [ a prov:Activity ;
          dcterms:title "JSI Newsfeed" ;
          prov:endedAtTime "2014-09-17T11:48:57+00:00"^^xsd:datetime ;
          prov:startedAtTime "2014-09-17T13:48:53.874994+02:00"^^xsd:datetime ;
          prov:wasAttributedTo <http://ijs.si> ] .
}

<http://ijs.si/enryched/211017736> {
  <http://ijs.si/article/211017736> a kdo:NewsArticle ;
      dcterms:created "2014-09-17T11:48:40.543007"^^xsd:datetime ;
      dcterms:language "en" ;
      dcterms:publisher [ rdfs:label "San Francisco Chronicle" ] ;
      dcterms:source <http://www.sfgate.com/news/crime/article/Authorities-make-arrest-in-community-
hall-shooting-5761145.php> ;
      dcterms:spatial [ gn:name "San Francisco",
              "USA" ;
          geo:lat 3.77796e+01 ;
          geo:long -1.2242e+02 ] ;
      dcterms:title "Authorities make arrest in community hall shooting" ;
      sioc:content """Authorities make arrest in community hall shooting
ABBEVILLE, La. (AP) -- Vermilion Parish authorities have arrested a 20-year-old Loreauville man in the
Saturday shooting at a community hall that injured seven people. Sheriff Mike Couvillon tells The Advocate
(http://bit.ly/1ARYf7L ) Bryson Provost was booked into the Vermilion Parish Correctional Center on seven
counts of attempted first-degree murder. Couvillon says Provost was arrested Monday afternoon at his
home by Iberia Parish Sheriff's Office deputies.
The shooting occurred at the Woodmen of the World Hall in Abbeville. The sheriff says none of the injuries
to the seven victims was life threatening. Couvillon said detectives continue to look at others who might
have been involved.He says a lack of full cooperation from victims and witnesses thus far is hindering the
investigation.""" ;
      sioc:topic "Christianity",
        "Denominations",
        "Guns",
        "North_America",
        "Parishes",
        "Recreation",
        "Religion_and_Spirituality",
        "Society",
        "United_States" ;
}
```

2 xLiMe's Approach for Statistical Content Linking

In this section, we explore the problem of cross-modal recommendation by analyzing the content generated from different modalities. We divide the section based on two different approaches.

- Using only the text generated from the modalities.
- Using the raw heterogeneous features extracted mainly from images and text.

2.1 Joint Text Space (JTS)

In this approach, we project every modality into a joint common space of a single modality. Here, we choose textual modality as the common representation of every multimedia item. Once every multimedia item is projected to text, JTS is then used to analyze modalities based on their textual content similarity.

The approach is dependent on the data generated by each of the content providers. For example, the input from the audio is dependent on the transcription of "speech to text" service of the video stream. While, the output of the image frames of video is dependent on the OCR recognition and automatic image annotation.

Similarly, data generated from social media may be present in the form of text. However, most of the social media are noisy. It lacks natural language structure and can be vague. We pre-process the data to improve the quality of the text extracted from it. Electronic programme guide (EPG) of TV shows is also given in the form of text. It is used for providing recommendations for similar TV shows.

Figure 5 shows the overall visualization of the approach.

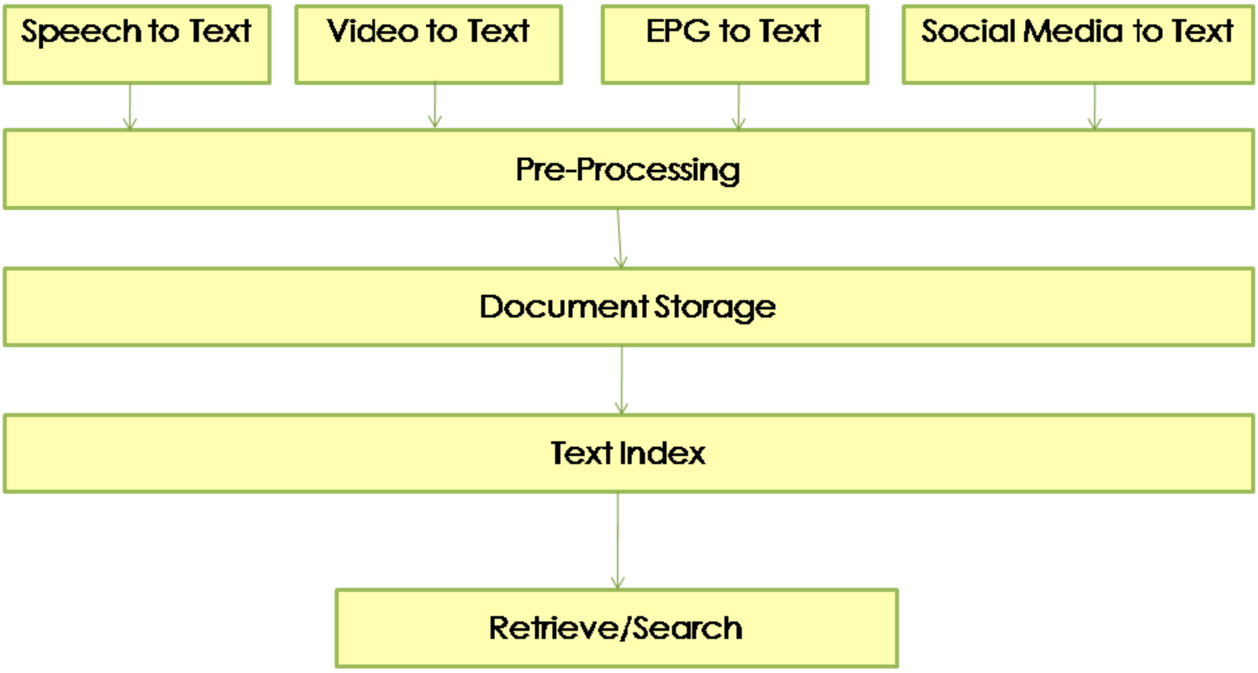


Figure 5: Overview of Joint Text Space Approach

In Figure 5, we observe that every modality transformed into text undergoes different processing steps. In the following sections we discuss each of these steps briefly.

2.1.1 Pre-Processing

The stream of data generated from different content providers is represented in the RDF format supporting the xLiMe Meta Data Model. The aim of the xLiMe Meta Data Model is to build a unified representation for all modalities to provide further functionalities such as semantic search.

JTS approach is built on the idea of using plain text. We extract only those important fields that are used for textual content similarity. To achieve it, we pre-process the data. Initially, the RDF format data generated from various content providers are converted to JSON representation. JSON is JavaScript object notation often used for lightweight data interchange. We use Apache Jena framework [7] to convert RDF format to JSON. Once the data format is interchanged, we extract important and relevant fields used for content similarity.

Most of the text obtained from tweets, forums and other social media sites contain noise in the form of junk characters, extra white spaces and line breaks. We clean the textual data by removing extra spaces, repeated line breaks, junk characters and non-ASCII or UTF-8 formats. Also, we convert the text to UTF-8 encoding if it was represented in ASCII format.

Since, every modality contains various fields. We pre-define the fields that will be used for storage. Below, we can see the templates that are used for each data stream in JSON format. This helps to push the data into standard document store for further processing and analysis.

VICO Social Media Template

{ u'Date':u' ', u'Publisher': u' ', u’Lang’: u’ ’,u'SourceURL': u' ', u'Text': u‘ ',u'_id': ObjectId('')}

Fields - Description

Date- Date on which article was published.

Publisher – Twitter, Facebook or Blog.

Lang – Language of the article

SourceURL – Source of the article.

Text – Text of the article.

Id – Object ID of the document store.

ZATTOO EPG Template

{ u'Date':u' ', u'Image Source': u' ', u'SourceURL': u' ', u'Text': u‘ ', u’Title’: u’ ‘,u'_id': ObjectId('')}

Fields - Description

Date- Date on which show is telecasted.

Image Source – Snapshot image of the show.

SourceURL – Source of the show in ZATTOO IPTV.

Text – Text of the article.

Title – Title of the show.

Id – Object ID of the document store.

JSI News Articles

{ u'Date':u' ', u’Lang’: u’ ‘, u'SourceURL': u' ', u'Title': u‘ ', u’Text’: u’ ‘,u'_id': ObjectId('')}

Fields - Description

Date- Date on which news article was published.

Lang – Language of the news article.

SourceURL – Source of the news article.

Text – Text of the news article.

Title – Title of the news article.

Id – Object ID of the document store.

In the following section, we see how the data represented in the templates are stored.

2.1.2 Document Store

Once the data is pre-processed, it is stored in a repository for future recommendations. We store the processed data in a schema-less document store called MongoDB [8]. This is a NoSQL database which supports JSON- like documents.

Below, we can see the sample JSON documents generated after pre-processing the VICO social media, EPG data and JSI news articles. We also added the document ids generated for each of these JSON documents after storing each of them in their respective collections of the MongoDB store.

VICO Social Media Data

```
{u'Date':u'2014-08-25 22:24:00',
 u'Publisher': u'Article',

U’Lang’:’de’,
 u'SourceURL': u'http://www.dasgelbeforum.de.org/board_entry.php?id=321187#1',
 u'Text': u‘Large Text',
 u'_id': ObjectId('5411a2200105f429a77dad72')}
```

ZATTOO EPG Data

```
{u'Date': u'2014-09-07 00:00:00',
 u'ImageSource': u'http://thumbod.zattoo.com/bbc-world-service/1410048060/8d24f81e/240x180.jpg',
 u'SourceURL': u'http://zattoo.com/program/14942468',
 u'Text': u'The latest international news from the BBC.',
 u'Title': u'BBC World News',
 u'_id': ObjectId('5411b0030105f42d0f4d2f75')}
```

JSI News Articles

```
{u'Lang': u'en',
u'SourceURL':u'http://www.bennadel.com/blog/2692-you-have-to-explicitly-end-streams-after-pipes-
break-in-node-js.htm',
u'Title': u'You Have To Explicitly End Streams After Pipes Break In Node.js',
u'Text': u'Large Text ',
u'Date': u'2014-10-07 14:03:42',
 u'_id': ObjectId('5433fbcbf0d32f19b2474371')}
```

There are numerous advantages of MongoDB. This is a document-oriented store which supports ad-hoc queries. It can also be used for indexing and has high availability. MongoDB can be scaled horizontally using sharding. This is a method for storing the data across machines for high throughput. The data divided across machines is called shards and is considered as an independent database. One of greatest advantage of sharding is that it reduces the amount of data that each server has to store. For instance the 256GB data can be stored in 4 shards each having 64GB of data.

MongoDB also supports any front-end programming languages for efficient visualization of the stored data. Once the data are stored in the repository, we index the text for fast retrieval and efficient recommendations.

In the following section we see how data stored in MongoDB are indexed.

2.1.3 Indexing

Once the documents are stored in MongoDB, documents are indexed based on their text fields. It is observed from the JSON samples of different content providers that there is more than one text field to index. For example the social media data are indexed based on the “Text” field while, the EPG data Index is created using the TV show “Title” and its “Description”. Similarly, news articles are indexed based on the “Title” and “Text” fields of the document.

To perform indexing, we use the distinct collections inside MongoDB store. Collections are logical partitions of the documents generated from various content streams stored inside a single MongoDB store. Each collection is indexed separately to support properties of the every content stream.

To achieve this, we use the “Indexing” feature of MongoDB. There are various options for indexing; one can create compound indexes, multi-key indexes, geo-spatial indexes, hashed indexes or text indexes.

We use the “Text indexing” feature to index multiple text fields of documents stored in distinct collections. The motivation for using the text indexing feature is its ability to search strings inside the documents. Below, we can see the approach used to index each collection inside MongoDB.

VICO Social Media Collection

```
db.socialmedia.ensureIndex( {Text: "text"}, {unique: true, dropDups: true} )
```

db – MongoDB store

socialmedia – Collection

ensureIndex – Creates a new Index, if there is no index.

- **Text** - Field to Index
- **Unique** – creates a unique index
- **dropDups** – Delete any duplicates in the index

ZATTOO EPG Data Collection

```
db.tvmetadata.ensureIndex( {Text: "text", Title: "text"}, {unique: true, dropDups: true} )
```

db – MongoDB store

tvmetadata – Collection

ensureIndex – Creates a new Index, if there is no index.

- **Text** - Field to Index
- **Title** - Field to Index
- **Unique** – creates a unique index
- **dropDups** – Delete any duplicates in the index

JSI News Articles Collection

```
db.jsinewsarticles.ensureIndex( {Text: "text", Title: "text"}, {unique: true, dropDups: true} )
```

db – MongoDB store

jsinewsarticles – Collection

ensureIndex – Creates a new Index, if there is no index.

- **Text** - Field to Index
- **Title** - Field to Index
- **Unique** – creates a unique index
- **dropDups** – Delete any duplicates in the index

It can be observed from the examples above that we used the extra parameters “Unique” and “dropDups”. As most of the data can be redundant, we created unique indexes and also removed duplicates.

In the future, we also aim to give importance to the text fields by weighing some text fields higher than others. For example, finding any keyword in the “Title” of a TV show can have more value than its description. This helps in achieving better content similarity.

Weights to the fields can be assigned during indexing and also during retrieval.

2.1.4 Content Similarity

Content similarity is used to provide cross-modal recommendations of social media and news articles for the ZATTOO TV video streams. We achieve this by finding content similarity between the text generated from the audio of TV video streams with social media data and News articles.

Below, we can see the sample text generated from the audio of ZATTOO video stream using speech to text transcription.

ZATTOO Audio transcription

xlime:hasASRText " not earnings and revenue relative solid follow his first post fact stares at the end of the BMW Championship survivor legislators around the city itself which went into my hands and no one time Yuvraj joining me on CNBC needs website gained exclusive access to some of the world's most fascinating people what a series of loneliness welcome back to top of the big corp " ;

Since most of the times the text generated from speech transcription can be noisy, we try to design an approach which is invariant to the noise in the text. We optimize text into keyword queries using a rapid automatic keyword extraction algorithm (RAKE) algorithm. This helps in searching the similar content stored in the indexed collections of social media and news articles.

Below we can see the output generated by the RAKE algorithm.

Text to Queries

```
[('revenue relative solid follow', 16.0),
('website gained exclusive access', 16.0),
('bmw championship survivor legislators', 16.0),
('post fact stares', 9.0),
('time yuvraj joining', 9.0),
('fascinating people', 4.0),
('big corp', 4.0),
('city', 1.0), ('cnbc', 1.0), ('end', 1.0), ('series', 1.0), ('loneliness', 1.0), ('back', 1.0), ('earnings', 1.0), ('hands',
1.0), ('world', 1.0), ('top', 1.0)]
```

RAKE is an efficient key-phrase extraction algorithm developed by [11]. It is independent of the domain of text and has only dependency on the stop words of a language. RAKE extracts the candidate keywords by splitting the document at stop words. It then assigns a score to each candidate keywords based on its co-occurrence with other words. The score for each candidate keyword is calculated based on its sum of individual words. Once the scores for each candidate keyword is calculated, top ranked keywords were considered as the input for matching documents.

We use the top ranked key-phrase to identify the relevant articles present in social media documents and News articles using a retrieval approach. The social media links and news articles retrieved is used to provide recommendations for the live TV shows telecasted.

2.1.5 Cross-modal Recommendations

In this section, we introduce the template used for recommendation and see an example of cross-modal recommendations obtained for the ZATTOO TV shows. Cross-modal recommendations are provided in JSON format satisfying the following template listed below.

Cross-Modal Recommendation Template

```
{ “tvmetadatarec”: “”, “cid”: “” , “jsinewsrec”: “”, “zattooid”: “”, “socialmediarec”: “ ”, “streamposition”:”” }
```

Tvmetadatarec - Recommendations of ZATTOO TV shows (List of URLs)

cid - ZATTOO Channel name (Text)

zattooid - ID of the show as stored by ZATTOO (Integer)

socialmediarec - Recommendations from VICO social media data (List of URLs)

jsinewsrec - Recommendations of JSI news articles. (List of URLs)

Streamposition - ZATTOO Image frame or Video stream position (Double)

Below, we provide a sample recommendations obtained from social media data, TV shows and news articles for the ZATTOO video stream position “8.490614e+07” aired on SKY news international.

Example 1

{"tvmetadatarec":

["http://zattoo.com/program/15216633", "http://zattoo.com/program/15156376",
"http://zattoo.com/program/15168144", "http://zattoo.com/program/15166900",
"http://zattoo.com/program/15153974", "http://zattoo.com/program/15155219",
"http://zattoo.com/program/15180855"],

"cid": "skynews-intl",

"jsinewsrec":

["http://www.chroniclelive.co.uk/news/local-news/north-east-news-replay-breaking-7920555",
"http://grantland.com/the-triangle/2014-mlb-playoffs-nlds-preview-cardinals-dodgers-giants-nationals/",
"http://www.oxfordtimes.co.uk/news/opinions/blogs/11508229.Parky_at_the_Pictures__DVD_2_10_2014
_/?ref=rss", "http://www.mouseplanet.com/10819/Animation_Anecdotes",
"http://www.nytimes.com/2014/10/05/fashion/alan-cumming-life-isnt-always-a-cabaret-Not-My-Fathers-
Son-memoir.html?partner=rss&emc=rss&_r=0",
"http://www.rockmnation.com/2014/10/7/6935627/missouri-offense-grades-maty-mauk-bud-sasser",
"http://www.carandsuv.co.nz/articles/target-acquired-308",
"http://www.bookslut.com/blog/archives/2014_10.php#020928", "http://www.nydailynews.com/life
style/west-young-fam-article
1.1973168?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%253A%2Bnydnrss%252Fg
ossip%252Fgatecrasher%2B%2528Gossip%252FConfidenti%2540l%2529",
"http://windycitymediagroup.com/lgbt/Bennett-and-Gagas-CD-Melissa-on-marriage-Laverne-in-
Vegas/49198.html", "http://www.liverpoolecho.co.uk/news/liverpool-news/re-read-hillsborough-inquests-
--monday-7923090", "http://www.mirror.co.uk/news/world-news/shrien-dewani-trial-recap-updates-
4428800", "http://grantland.com/the-triangle/mlb-the-30-2014-season-finale/",
"http://www.haskell.org/haskellwiki/index.php?title=Xmonad%2FFrequently_asked_questions&diff=58959
&oldid=58341", "http://orthodoxwiki.org/index.php?title=User:Flux&diff=120011&oldid=119966",
"http://www.emirates247.com/sports/formula-1-jules-bianchi-has-traumatic-brain-injury-family-2014-10-
05-1.565211", "https://adactio.com/articles/6574", "http://westhawaiitoday.com/opinion/letters/letters-
10-7-14", "http://www.rcgroups.com/forums/showthread.php?t=2255313",
"http://coloradopols.com/diary/63699/coffman-v-romanoff-round-4-live-enough-blog",
"http://www.kunc.org/post/whats-it-be-neil-patrick-harris-he-gives-you-options",
"http://www.grandprix.com/race/r912friquotes.html", "http://www.automotiveaddicts.com/47721/2015-
cadillac-escalade-esv-4wd-premium-review-test-drive", "http://www.jamesaltucher.com/2014/10/life-is-
like-a-game-heres-how-you-master-any-game/", "https://adactio.com/articles/6546"],

"zattooid": "15433701",

"socialmediarec": ["http://pepolino.eu/2010/08/07/piratenschuhe/comment-page-35/", "http://m-
maenner.de/2014/05/miteinander-kaempfen/comment-page-78/",
"http://www.vwaudiforum.co.uk/forum/showthread.php?161174-Early-airing-for-all-new-Audi-TT-
Roadster-ahead-of-Paris-show-world-debut#1",
"http://x3.xbimmers.com/forums/showthread.php?t=408177&page=7#138", "http://m-
maenner.de/2014/05/miteinander-kaempfen/comment-page-76/", "http://www.kampfkunst-
board.info/forum/f18/flachwitze-22194/index317.html#4759", "http://m-
maenner.de/2014/05/miteinander-kaempfen/comment-page-84/", "http://blog.ruralvive.com/inturferia-
turismo-interior-valladolid/comment-page-18/",
"http://forum.skyscraperpage.com/showthread.php?t=177316&page=338#6743", "http://www.android-
hilfe.de/5766113-post1.html", "http://forum.notebookreview.com/samsung/733030-2013-ativ-book-9-

```
plus-owner-s-lounge-np940x3g-163.html#1624",
"http://trabajoporinternetcolombia.blogspot.com/2014/09/tienda-virtual-wwwavenspartancom-en.html",
"http://anonymous-die-hetzer.blogspot.com/2014/06/ubersicht-fur-we-are_9119.html",
"http://isialada.blogspot.com/2014/09/manual-sobre-experiencias-fuera-del.html", "http://m-
maenner.de/2014/05/miteinander-kaempfen/comment-page-72/", "http://m-
maenner.de/2014/05/miteinander-kaempfen/comment-page-85/", "http://m-
maenner.de/2014/05/miteinander-kaempfen/comment-page-99/", "http://www.trojaner-
board.de/159242-windows-xp-email-zip-attachment-danach-dateien-verschluesselt.html#4",
"http://aequitas-jmp-blogspotcom.blogspot.com/2014/08/i-expedicion-38-expedition-38-eei-una.html",
"http://nettipaivakirja5.blogspot.com/2014/09/putins-doomsday-plane-circling-near.html",
"http://pepolino.eu/2010/08/07/blumchenschuhe/comment-page-35/",
"http://www.alternatehistory.net/Discussion/showthread.php?t=315901#12", "http://m-
maenner.de/2014/05/miteinander-kaempfen/comment-page-68/", "http://m-
maenner.de/2014/05/miteinander-kaempfen/comment-page-94/",
"http://forums.hardwarezone.com.sg/eat-drink-man-woman-16/running-man-fans-who-yr-fav-running-
man-part-9-a-4767620-244.html#3654"],

"streamposition": "8.490614e+07"}
```

2.2 Joint Feature Space

In the previous section, we have seen an approach using only the text generated from different modalities. In this section, we conduct research to achieve content similarity based on the raw features generated from different modalities. In the following sections, we formulate the problem of cross-modal content similarity as a retrieval problem and present our approach.

2.2.1 Architecture

ZATTOO video streams are used to find cross-modal recommendations from social media data and news articles. In the previous approach, we have used the audio of the video streams to get recommendations.

In this approach, we analyze the research question of using video as a collection of image frames. We assume each image frame will be accompanied by other modalities like text from speech of the video and subtitles in the form of text. Assuming each image frame and the text belong to some particular semantic category, we design a following approach.

Figure 6 depicts the approach.

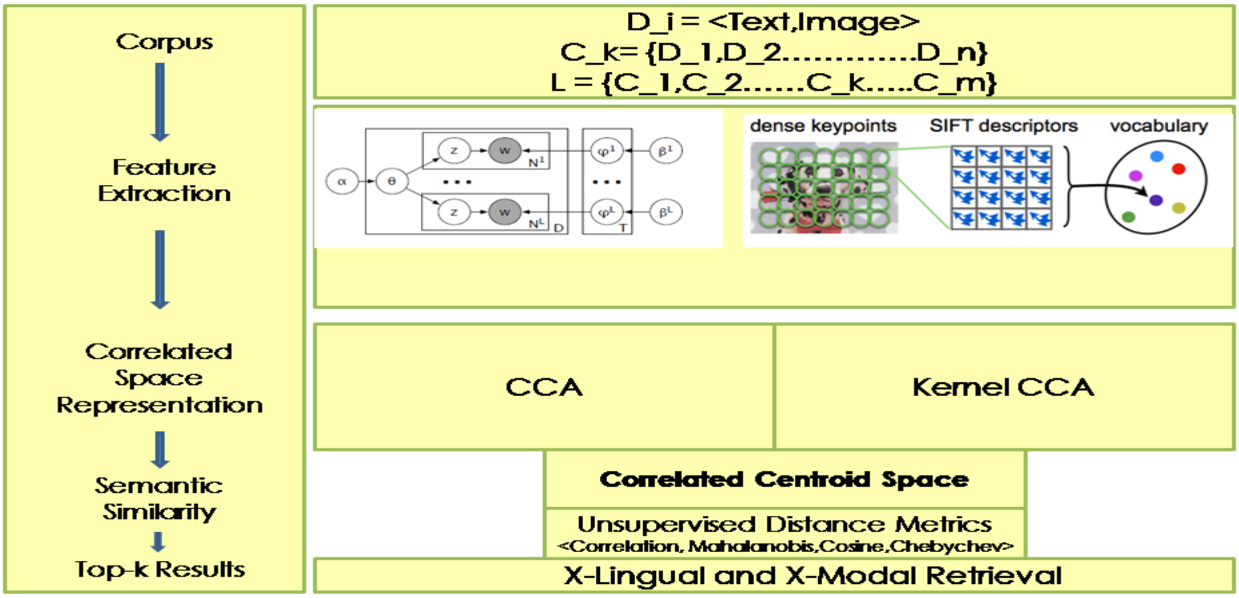

Corpus
Feature Extraction
Correlated Space Representation
Semantic Similarity
Top-k Results
D_i = <Text,Image>
C_k= {D_1,D_2............D_n}
L = {C_1,C_2......C_k.....C_m}
dense keypoints
SIFT descriptors
vocabulary
CCA
Kernel CCA
Correlated Centroid Space
Unsupervised Distance Metrics
<Correlation, Mahalanobis,Cosine,Chebychev>
X-Lingual and X-Modal Retrieval

Figure 6: The xLiMe Approach using heterogeneous features

In Figure 6, we can see two different columns showing the flow of the approach. We divide the entire approach into four different steps. Each step is performed on two different modalities of data (e.g. images and text).

The four steps used to provide cross-modal similarity are as follows.

- Extraction of raw features from images and text.
- Correlated or Joint space representation of raw features extracted from text and images.
- Modification of the correlated space using clustering.
- Vector based similarity measures for finding cross-modal similarity
- Ranking of similar content

Below, we describe each step briefly.

2.2.2 Feature Extraction

Once the data are represented in the form of image and text modalities, we extract the features that represent the content. Below, we describe the approaches used to extract features from both text and images.

2.2.2.1 Text Features

Polylingual Topic Models (PTM) [12] is used to extract features from the text in the form of topics distribution. We choose PTM for feature extraction as it supports multiple languages and generates same topics across languages. Around 250,000 Wikipedia articles aligned to each other in English, German and Spanish is used to learn PTM.

We trained 10 topics on this Wikipedia corpus for each of English, German and Spanish languages.

2.2.2.2 Image Features

SIFT [13] local feature vectors are extracted from images. All these local feature vectors are kept in a single set. The K-Means clustering algorithm is applied over the set of local feature vectors to find the centroid coordinates. This set of centroids is the final vocabulary. The global feature vector is a histogram that counts how many times each centroid has occurred in an image.

We use the nearest centroid for each local feature to construct the histogram. The final output is a 128-dimension feature vector for each image in the collection.

2.2.3 Correlated Space

We use canonical correlation analysis (CCA) [14] to learn the subspace that maximizes the correlation between heterogeneous features of images and text. Equation 1 shows the maximization procedure between the text and image modalities.

$$\arg\max_{u,v} \frac{u' \Sigma_{ti} v}{\sqrt{u' \Sigma_{tt} u}\sqrt{v' \Sigma_{ii} v}} ----------------(1)$$

The u and v matrices represent the text and image projections obtained after maximizing the correlations between image and text modalities respectively. While $\sum_{tt}$ represents the covariance matrix of the text modality, $\sum_{ii}$ represents the covariance matrix of the image modality. $\sum_{ti}$ is the cross-covariance matrix between the text and image modalities. Equation 1 is solved using a generalized eigen value problem.

The final output of CCA provides the joint dimensionality reduction on both image and text features and output the vectors that maximises the correlation between them.

Kernel Version of CCA can be used to find the correlation between non-linear relationships. We now replace the input matrices with their kernel versions and maximize the correlation. Equation 2 shows the kernel version of CCA.

$$\arg\max_{x,y} \frac{x' K_T K_I y}{\sqrt{x' K_T^2 x}\sqrt{x' K_I^2 y}} ---------------(2)$$

The x and y matrices represent the text and image projections obtained after maximizing the correlations between image and text modalities respectively. K_T and K_I represent the positive semi-definite kernel matrices of text and image modalities respectively.

2.2.4 Correlated Centroid Space

The vector spaces obtained after kernel CCA of the input text and image modalities is used to find the centroids. We achieve this by performing two steps.

1) Assignment Step – Assigns the each observed sample in the transformed feature space of image and text data to its closest mean.
2) Update Step – Calculate the new means that will be the centroid.

Once the k-centroids are obtained for the k-classes in the data, we modify the correlated space of each sample by replacing its features space with its closest centroids. Now the modified feature space is used to find the semantic similarity between the new samples.

2.2.5 Semantic Similarity

The semantic similarity between the cross-modal documents is found using different unsupervised similarity measures. The features extracted from the text and images represented using the correlated centroid space are used to find the similarity.

We evaluated few similarity measures like correlation distance, cosine similarity, Mahalanobis distance and Chebychev distance and chose the best one which maximizes the semantic similarity.

3 Evaluation

In this section, we present the results obtained on the wiki dataset [9] using the joint feature space approach.

3.1 Dataset

We used the wiki dataset [9] created for English texts and images using the Wikipedia featured articles. The dataset consists of 2173 training and 693 testing documents belonging to 10 different semantic categories taken from art, biology, sport etc. We expanded the dataset into two more languages i.e. German and Spanish while keeping the original images intact. The final dataset consists of three sets of image and text pairs.

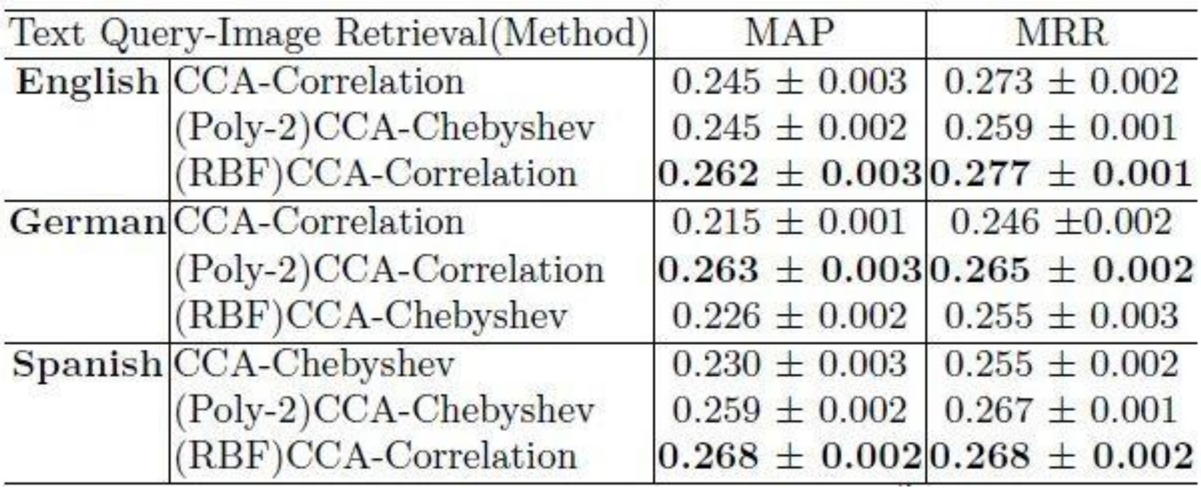
To evaluate the joint feature space approach, we used mean average precision (MAP) and mean reciprocal rank (MRR). A high value of MAP and MRR shows the effectiveness of the approach in analyzing cross-modal content in different languages.

3.2 Experiments

Below, we performed two different experiments. First, using an image query to find the relevant text articles and second using a text query to find the relevant image articles.

The tables below show the results obtained for English, German and Spanish using CCA and kernel versions of CCA i.e. Polynomial with degree 2(poly-2) and RBF with the best distance metric that maximise the semantic similarity based on MAP scores.

Text Query – Image Retrieval



Text Query-Image Retrieval(Method)		MAP	MRR
English	CCA-Correlation	0.245 ± 0.003	0.273 ± 0.002
	(Poly-2)CCA-Chebyshev	0.245 ± 0.002	0.259 ± 0.001
	(RBF)CCA-Correlation	**0.262 ± 0.003**	**0.277 ± 0.001**
German	CCA-Correlation	0.215 ± 0.001	0.246 ±0.002
	(Poly-2)CCA-Correlation	**0.263 ± 0.003**	**0.265 ± 0.002**
	(RBF)CCA-Chebyshev	0.226 ± 0.002	0.255 ± 0.003
Spanish	CCA-Chebyshev	0.230 ± 0.003	0.255 ± 0.002
	(Poly-2)CCA-Chebyshev	0.259 ± 0.002	0.267 ± 0.001
	(RBF)CCA-Correlation	**0.268 ± 0.002**	**0.268 ± 0.002**

Image Query – Text Retrieval

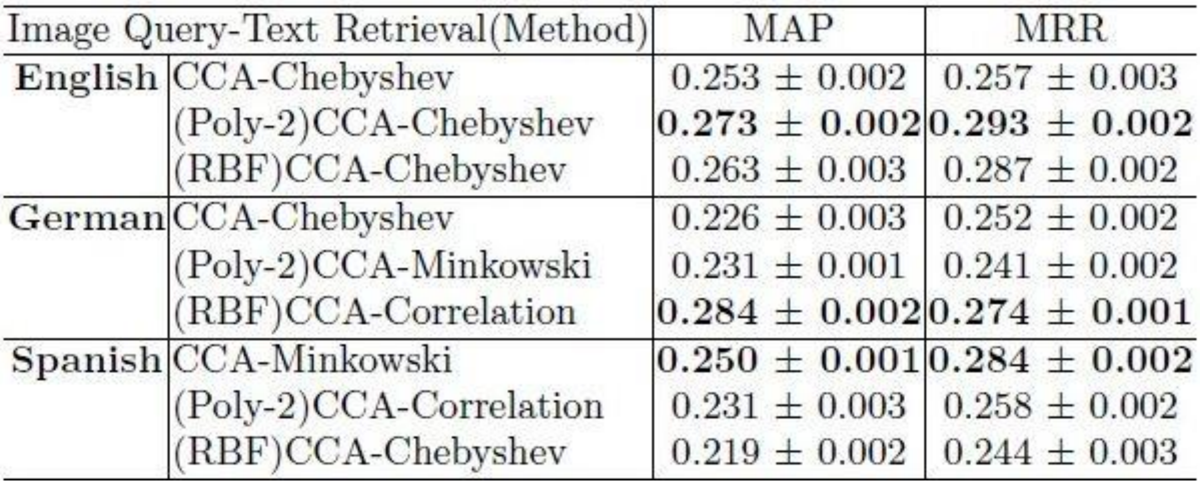


Image Query-Text Retrieval(Method)		MAP	MRR
English	CCA-Chebyshev	0.253 ± 0.002	0.257 ± 0.003
	(Poly-2)CCA-Chebyshev	**0.273 ± 0.002**	**0.293 ± 0.002**
	(RBF)CCA-Chebyshev	0.263 ± 0.003	0.287 ± 0.002
German	CCA-Chebyshev	0.226 ± 0.003	0.252 ± 0.002
	(Poly-2)CCA-Minkowski	0.231 ± 0.001	0.241 ± 0.002
	(RBF)CCA-Correlation	**0.284 ± 0.002**	**0.274 ± 0.001**
Spanish	CCA-Minkowski	**0.250 ± 0.001**	**0.284 ± 0.002**
	(Poly-2)CCA-Correlation	0.231 ± 0.003	0.258 ± 0.002
	(RBF)CCA-Chebyshev	0.219 ± 0.002	0.244 ± 0.003

3.3 Result Analysis

It can be observed from both tables that kernel versions of CCA outperformed the baseline CCA on MAP scores. This shows the presence of non-linearity in the data and the efficiency of joint feature space approach in handling it.

Also, the approach is scalable and is dependent only the features extracted from the multimedia data. When the dimension of text features is increased, the dimension of the image features has also to be increased for allowing an effective joint dimensionality reduction.

4 Conclusion

In this deliverable, we presented statistical content linking approaches for analyzing the cross-modal documents. We proposed two approaches to solve this problem. First solution uses only the text generated by diverse modalities, while the second approach uses heterogeneous features extracted from each modality.

We observed that the joint text based approach is easily scalable, is fast and can provide better solutions if all modalities can be efficiently transformed to text. On the other hand, the joint feature space works without any dependency on the text and is also scalable.

In the future, we aim to perform a joint analysis of both approaches for efficiency and speed.

5 Acknowledgements

We would like to thank Inga Shamkhalov for her help in improving the presentation of the deliverable.

References

[1] http://thelocal.de

[2] http://flickr.com (Flickr)

[3] http://youtube.com (YouTube)

[4] http://www.zattoo.com (ZATTOO)

[5] http://www.vico-research.com/ (VICO)

[7] https://jena.apache.org/(Jena Framework)

[8] http://www.mongodb.org/ (MongoDB)

[9] http://www.svcl.ucsd.edu/projects/crossmodal/

[10] xLiMe Meta Data Model

[11] Automatic keyword extraction from individual documents (RAKE). Text Mining: Applications and Theory.

[12] Mimno, David, et al. Polylingual topic models. 2009. in Proceedings of the Empirical Methods in Natural Language Processing (EMNLP). Vol 2. ACL.

[13] Lowe, David G. Object recognition from local scale-invariant features. 1999. The proceedings of the seventh IEEE international conference on computer vision. Vol. 2.

[14] Relation between Two Sets of Variates. (CCA). 1936. Biometrika. Vol. 28, No. ¾.