chist-era

MUCKE

Multimedia and User Credibility Knowledge Extraction

http://ifs.tuwien.ac.at/~mucke/

Mihai Lupu

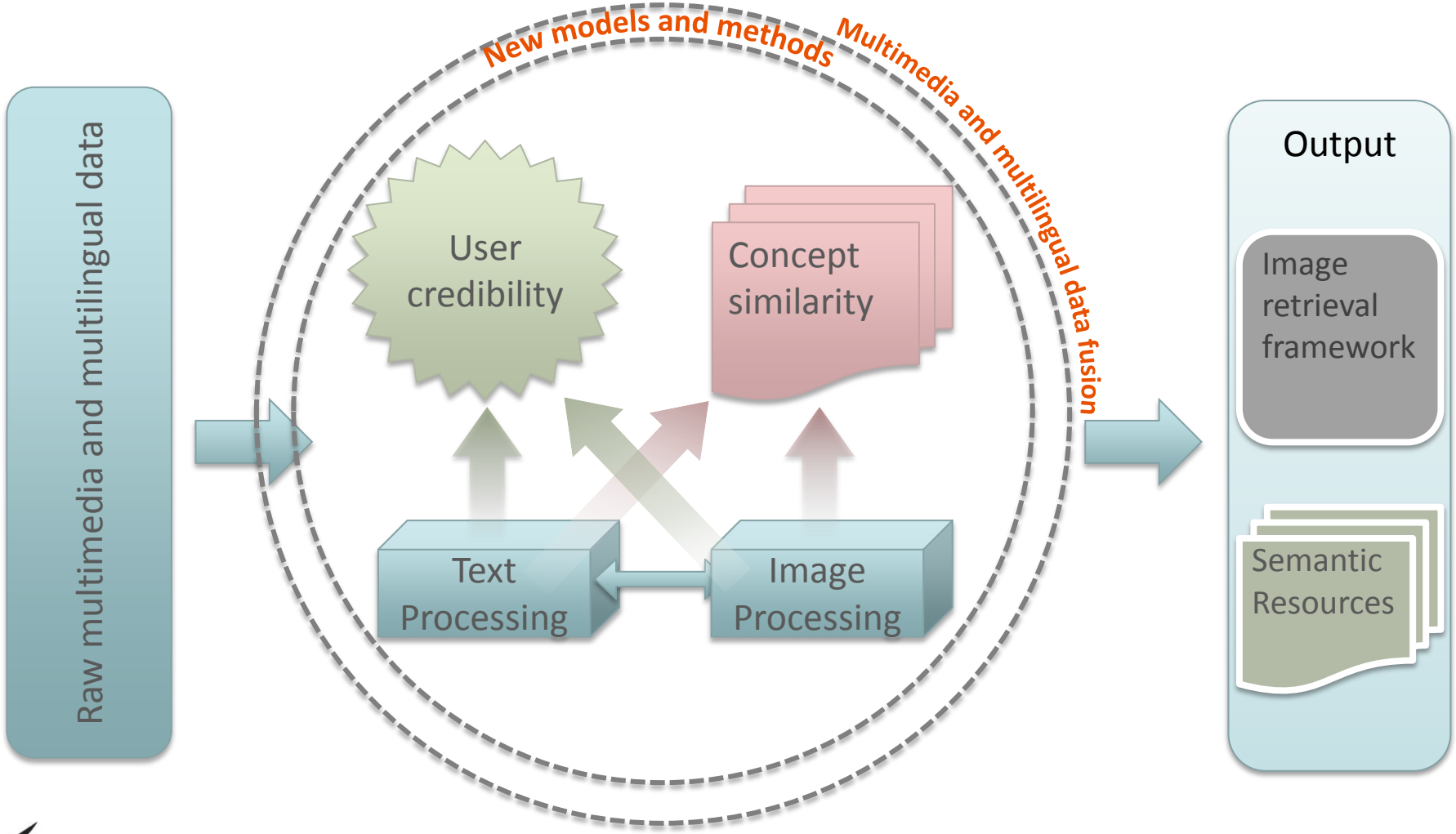
Vienna University of Technology

lupu@ifs.tuwien.ac.at

Project status

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22	M23	M24	M25	M26	M27	M28	M29	M30	M31	M32	M33	M34	M35	M36	Total
WP 1 - Data Collection	2.8	2.8	2.8	2.8	2.8	2.8	1.8	1.8	1.8	1.8	1.8	1.8																									**28**
T1.1 Assessment and Collection of Existing Resources	1.0	1.0	1.0	1.0	1.0	1.0																															6
T1.2 New Data Collection	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8																									22
WP 2 - Multimedia Processing and Fusion		2.8	2.8	2.8	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	1.4	1.4	1.4																**70**
T2.1 Text processing		1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2																			20
T2.2 Image processing		1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6																			27
T2.3 Multimedia fusion					1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4																23
WP 3 - User Credibility for Multimedia Streams						1.4	1.4	1.4	1.4	1.4	3.6	3.6	3.6	3.6	3.6	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2													**45**
T3.1 Credibility Model Definition						1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4																						14
T3.2 User Context Mining											1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0													14
T3.3 User content mining and fusion											1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2													17
WP 4 - Multimedia Concept Similarity Models													2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	**60**
T4.1 Textual Concept Similarity													1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	28
T4.2 Multimedia Concept Similarity													1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	32
WP5 - User-Centred Image Retrieval																								2.7	2.7	2.7	5.4	5.4	5.4	5.4	4.1	4.1	4.1	2.7	2.7	2.7	**50**
T5.1 - Image Search Results Diversification																								1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4				14
T5.2 - User-Centred Image Retrieval Model																								1.3	1.3	1.3	1.3	1.3	1.3	1.3							9
T5.3 - Retrieval Prototype Integration																											2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	27
WP6 - Evaluation and Resource Sharing						0.8	2.3	2.3	2.3	0.8	0.8	1.3	1.3	1.3	1.3	1.3	1.3	1.3	0.8	0.8	0.8	1.3	1.3	1.3	2.8	2.8	2.8	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	**44**
T6.1 - Algorithms and Prototype Evaluation												0.5	0.5	0.5	0.5	0.5	0.5	0.5				0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	11
T6.2 - Shared Task for User Credibility Estimation						0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	24
T6.3 - Resource Sharing							1.5	1.5	1.5																1.5	1.5	1.5										9
WP7 - Project Management	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	**5**
Total	3.0	5.7	5.7	5.7	7.1	9.3	9.8	9.8	9.8	8.3	10.5	11.0	11.6	11.6	11.6	10.2	10.2	10.2	7.0	7.0	7.0	6.1	6.1	8.8	8.1	8.1	10.8	9.3	9.3	9.3	8.0	8.0	8.0	6.6	6.6	6.6	**302**
																			D2.2						D3.3												D6.1
							D1.1			D6.3			D1.2			D3.1			D2.1			D2.3			D3.2			D4.2			D5.2			D5.1			D5.3
																									D4.1												D6.2

The core


New models and methods
Multimedia and multilingual data fusion
Raw multimedia and multilingual data
User credibility
Concept similarity
Text Processing
Image Processing
Output
Image retrieval framework
Semantic Resources

The data

Raw multimedia and multilingual data

- Existing collections
  - A survey done and published online
    - ImageNet – 14 million annotated images
    - mediaEval – 3.2 million images
    - MIRFLICKR – 1 million annotated images
    - Wikipedia (DBpedia)
    - ClueWeb09/12
- New data
  - Aim: 100million annotated images
  - Crawling ongoing

The data

Raw multimedia and multilingual data

- Distributed crawling and replicated storage
  - CEA – code for crawling
  - TUW – task distribution
  - All – collecting data

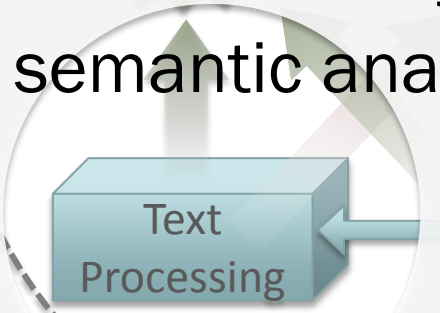
Output

Image retrieval framework

Semantic Resources

Text Processing

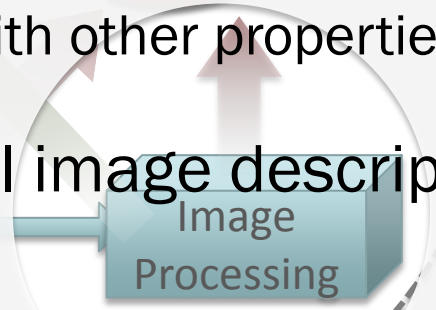
- Entity recognition
- Disambiguation
- Anaphora resolution
- Combined with IR methods
  - Latent semantic analysis
  - Explicit semantic analysis



- Components for:
  - English, French, German, Romanian

Image Processing

- Parsimonious image description
- Large scale concept detection
- Detector generalization
  - Across different datasets
- Asses the use and utility of
    - Different local image descriptors
    - their combination with other properties (e.g. color)
  - For optimal low-level image description



- Adapted models for specialized tasks
  - Face / landmark recognition

Concept detection

Large number of classifiers in a hierarchical manner

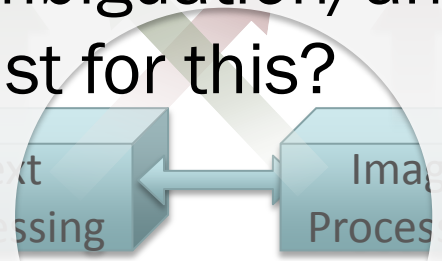
Low level : visual words
both for entire image and
for image patches(salient points, regular sampling or grids)
Features : Such as color, gradient histograms, GIST, etc

Mid level
Categories common in high level conceps
Attribute classifiers
Adjectives (color category, big/small, texture categories, photo related- dark/ bright
Parts : wheels, windows , etc

High level
Detector outputs
Photo/illustration
black and white/ color
İndoor / outdoor
Has a face / has a person / number of faces / size of faces
Objects : Animal / car / building / tree / sky / .... (Especially train part based classifiers for several simple concepts)

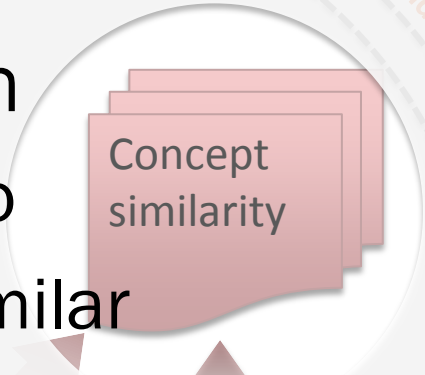
Multimedia fusion

- The “IR” way
  - Extracting probabilities of relevance from scoring functions
    - Score normalization
  - Can we get improvements from the text mining (disambiguation/anaphora resolution) just for this?



Concept similarity

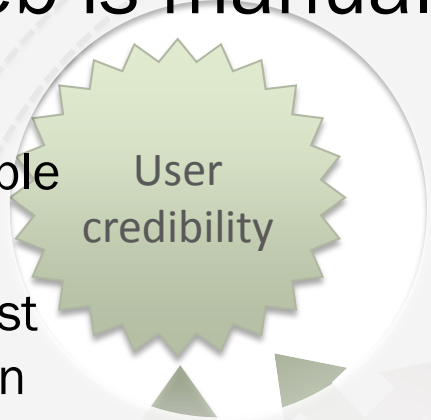
- Bring multimedia fusion to a higher level
- Research question
  - How similar are two concepts/what is similar when they are not even represented in the same modality?



- Investigation path
  - Link to an existing knowledge base
    - E.g. BabelNet 1.1 http://babelnet.org/

User Credibility

- Large amounts of information in the social web is manually created.



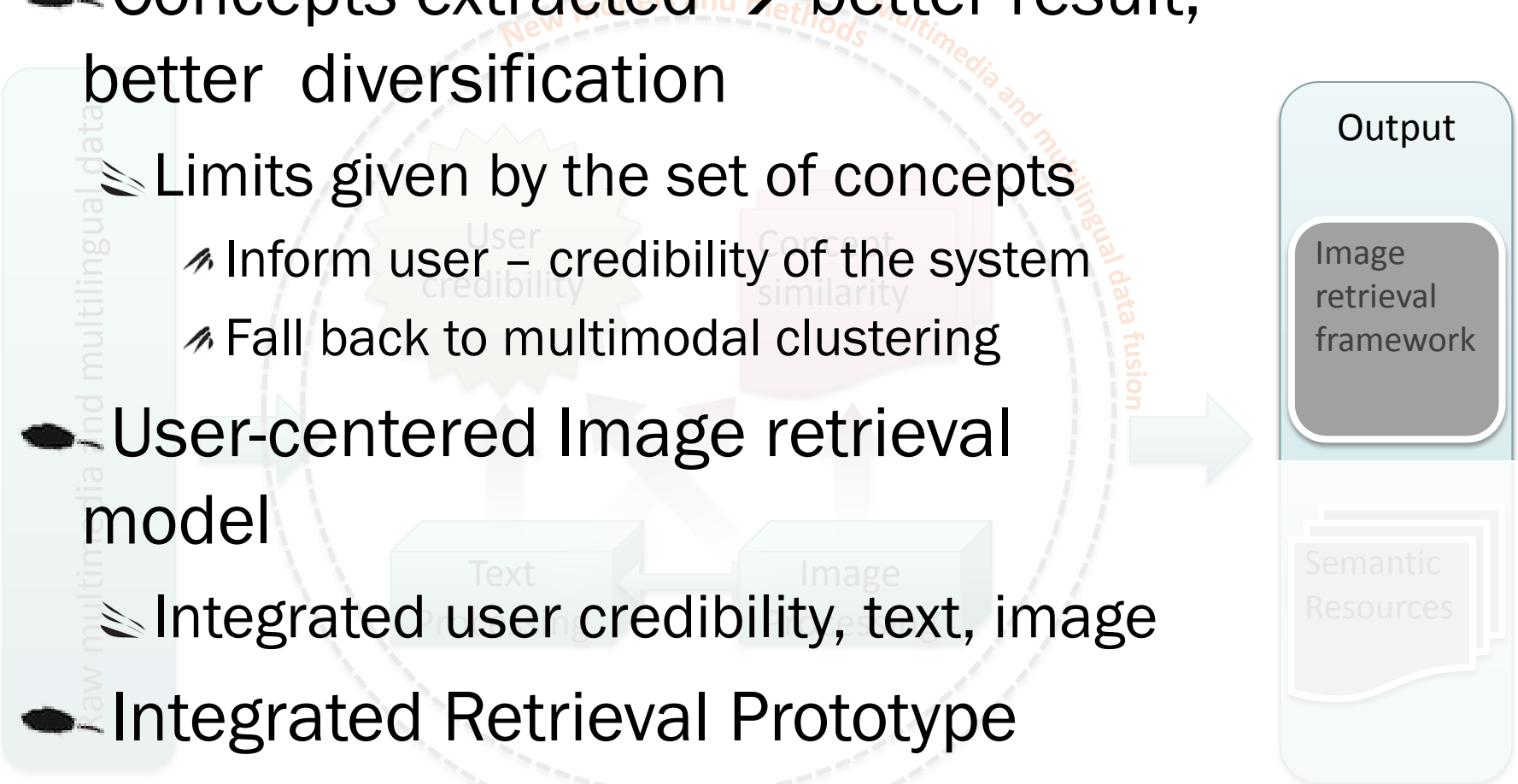
- Expertise
  - how knowledgeable is the user?
  - How much interest does [s]he have in the concept?
- Trustworthiness
  - Previous actions
  - Location in social network
- User context mining
- User content mining
  - Research questions:
    - How to put a number on this?
    - How is this integrated into the retrieval method?
  - Investigation path:
    - With manually created annotations + machine learning

Credibility

- Credibility is one central concept in information consumption & assessment
  - We read news from media we trust
  - We vote for politicians who will (hopefullly) keep their campaign promises
  - …
- Credibility for Web 1.0 (authored content)
  - Graph-based definition - links toward a site are equated as trust marks
  - Sites are ranked according to their position in the graph
  - Context based trust score
- Credibility for Web 2.0 (dynamic web)
  - ???
- Credibility for Web 3.0 (collaborative platforms)
  - User credibility (or expertise) present in many forums
  - Mostly derived through crowdsourcing – votes from other users
  - Existing work about automatic credibility estimation for Q&A sites
- Credibility for multimedia streams shared on social networks is poorly studied

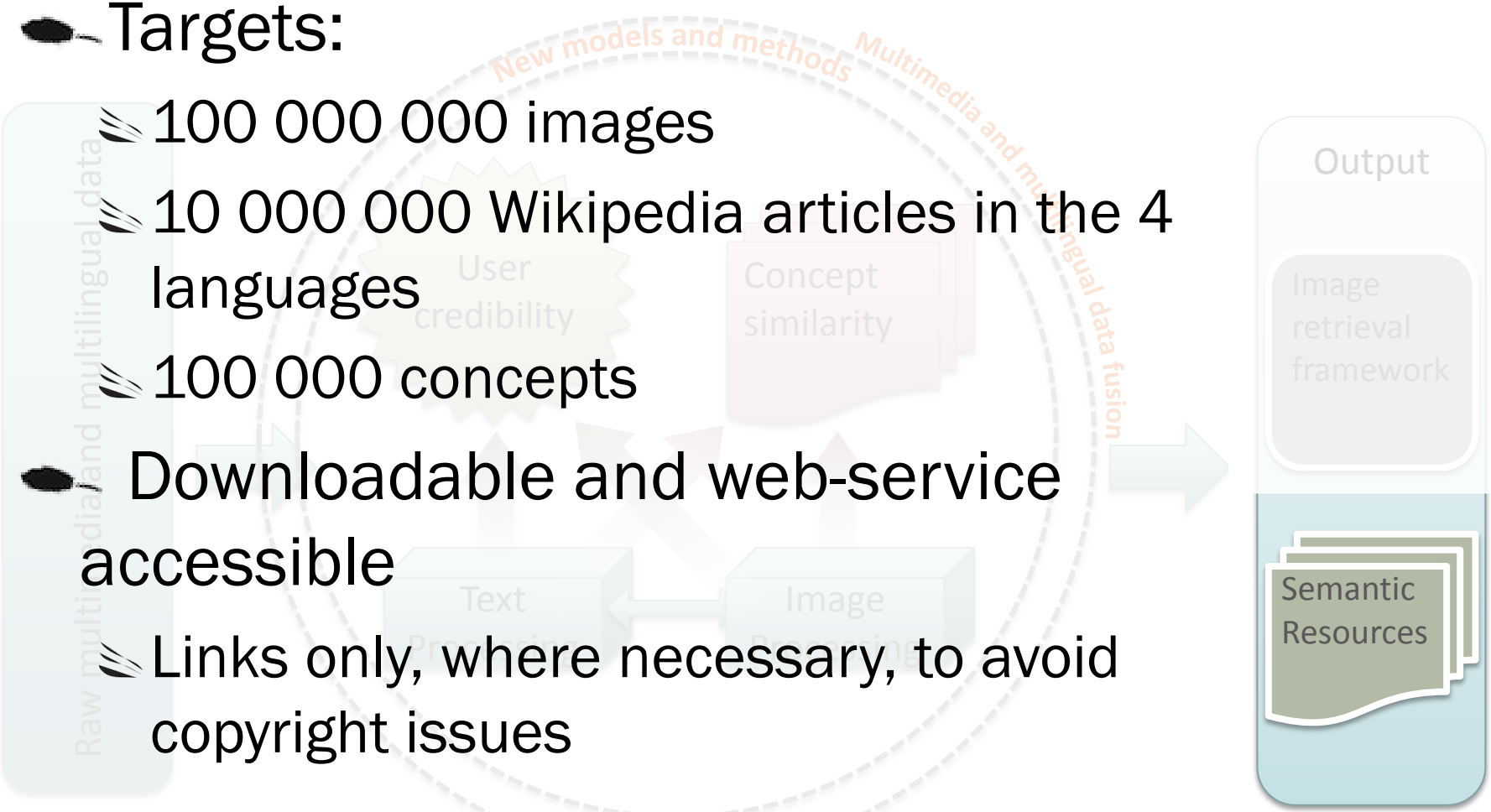
User-centred image retrieval

- Concepts extracted → better result, better diversification
  - Limits given by the set of concepts
    - Inform user – credibility of the system
    - Fall back to multimodal clustering
- User-centered Image retrieval model
  - Integrated user credibility, text, image
- Integrated Retrieval Prototype



Semantic resources

- Targets:
  - 100 000 000 images
  - 10 000 000 Wikipedia articles in the 4 languages
  - 100 000 concepts
- Downloadable and web-service accessible
  - Links only, where necessary, to avoid copyright issues



Evaluation

- Individual components in existing evaluation campaigns
  - QA4MRE, SEMEVAL
  - MediaEval, ImageCLEF, TRECVid
- Credibility task to be created
  - User ranking on credibility (for specific topics)
  - Given image+text+metadata – rank users based on their perceived credibility
    - Manually created assessments

Team

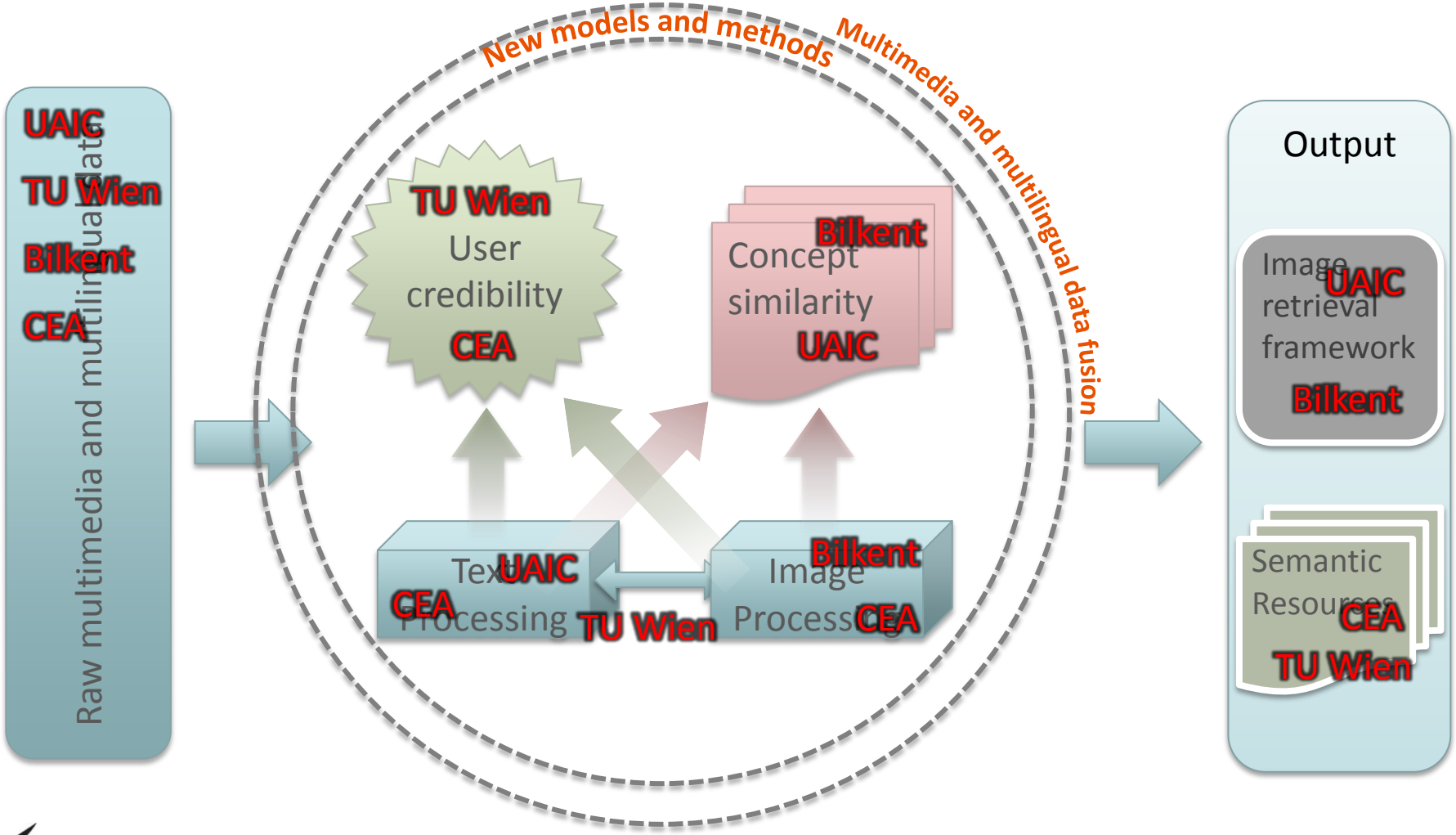
- Bilkent University, Turkey
- “Al. I. Cuza” University, Iasi, Romania
- Vienna University of Technology, Austria
- Center for Alternative and Atomic Energy, France

Team

- B
- “A a
- V tria
- O ergy,

Fr

The core


New models and methods
Multimedia and multilingual data fusion
Raw multimedia and multilingual data
UAIC
TU Wien
Bilkent
CEA
TU Wien
User credibility
CEA
Bilkent
Concept similarity
UAIC
Text Processing
UAIC
CEA
TU Wien
Image Processing
Bilkent
CEA
Output
Image retrieval framework
UAIC
Bilkent
Semantic Resources
CEA
TU Wien

CEA : LVIC - Laboratory for Vision and Content Engineering

- ~ 60 persons in all, with 25 people working on multimedia
  - 30 ongoing projects for the multimedia theme
    - Periplus, Egonomy, DataScale, ePoolice
    - Large number of direct collaborations with industrial partners
  - ~25 publications/year
- Objective – understand and describe multimedia documents (text, image, video)
  - Information retrieval over multimedia collections
  - Document filtering using domain related criteria
  - Document summarization and presentation
- Application domains
  - Electronic content Management
  - Cultural heritage and tourism applications
  - Collaborative filtering for product and service proposal
  - Technological watch
- Participation to/organization of evaluation campaigns

BILKENT University

- the first private, nonprofit university in Turkey
- founded on October 20, 1984
- “Bilkent” = an acronym of "bilim kenti": Turkish for "city of learning and science.”

- Computer Engineering Department
  - 22 faculty members
  - algorithms, artificial intelligence, bioinformatics, computer architecture, computer graphics, computer networks, computer vision, cryptography, data mining, database systems, information retrieval, machine learning, parallel and distributed systems, performance evaluation, scientific computing, and software engineering.

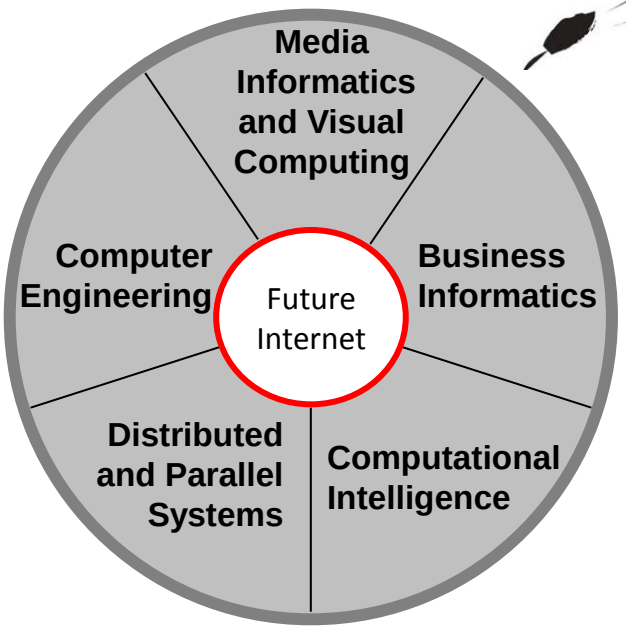
“Al. I. Cuza” University

- Computer Science Department
  - 20 years Faculty of Computer Science
  - ~ 1400 students (1150 Bachelor, 200 Master,50 PhD Students)
  - ~ 40 Professors (9 Full Professors)
- Research Projects
  - Natural Language Processing – Dan Cristea
  - Software Engineering – Dorel Lucanu

- NLP
  - METANET4U
  - ATLAS
  - LT4eL
  - ELIAS
  - eDTLR
  - CLEF, TAC, RTE campaigns

- multilingualism, services, resources

TU Wien



- Informatics Faculty
  - 18 Bachelor Programs
  - 43 Master Programs
  - Doctoral Program
  - 4.300 Employees
  - 25.000 Students
  - 150 Full Professors
  - 1.90 Scientific Personnel
- Information Management and Preservation Lab
  - Data Mining and Machine Learning
  - Information Retrieval
  - Digital Preservation
  - Led by Prof. Andreas Rauber
  - 20 people (of which 19 funded by external funds)

Project communication

- Internal
  - Meetings
    - Kick-off meeting - Vienna, 22-23 Nov 2012
    - S1 – Istanbul, 3-4 April 2013
  - Wiki
- External
  - Website http://ifs.tuwien.ac.at/~mucke
  - Publications
    - So far: 1 paper accepted, 1 submitted
  - Joint proposal for workshop on credibility at SIGIR 2013
  - Exchanges
    - TUWien – NII researcher exchange on credibility in information retrieval (June-July 2013)

Summary

- For the task of multimedia retrieval, MUCKE introduces new concepts and model - merge topical relevance and domain specific user credibility
- Using the yet untapped data in multimedia retrieval, social networks, and by creating semantic descriptors of groups and using them to calibrate probabilities of semantic tags applied to individual data → more relevant results
- Our transition from scores to probabilities allows the systems to be aware of low levels of confidence in their results
- A mixture fusion approach (applied late, but based on early processing), based on moving from ranking scores to probability values which can be applied to merge any type of data

chist-era

Thank you

MUCKE

Multimedia and User Credibility Knowledge Extraction

http://ifs.tuwien.ac.at/~mucke/