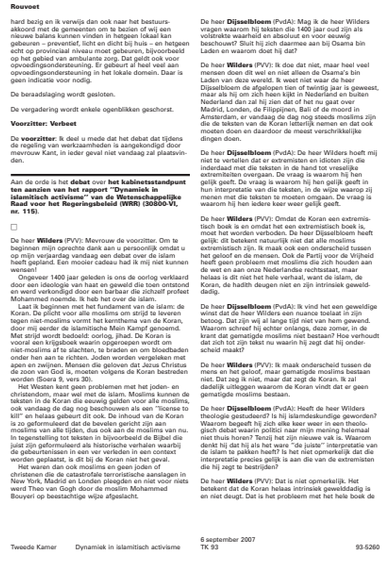
Indexing of Political Texts: Combining Different Sources of Evidence

Mostafa Dehghani, Hosein Azarbonyad, Maarten Marx, Jaap Kamps

UNIVERSITEIT VAN AMSTERDAM

Indexing Political Text

- Indexing political documents using thesauri



Rouvoet

hard bezig en ik verwija dan ook naar het bestuursakkoord met de gemeenten om te bezien of wij een nieuwe balans kunnen vinden in hetgaan lokaal kan gebeuren – preventief, licht en dicht bij huis – en hetgaan echt op provinciaal niveau moet gebeuren, bijvoorbeeld op het gebied van ambulante zorg. Dat geldt ook voor opvoedingsondersteuning. Er gebeurt al heel veel aan opvoedingsondersteuning in het lokale domein. Daar is geen indicatie voor nodig.

De beraadslaging wordt gesloten.

De vergadering wordt enkele ogenblikken geschorst.

Voorzitter: Verbeet

De **voorzitter**: Ik deel u mede dat het debat dat tijdens de regeling van werkzaamheden is aangekondigd door mevrouw Kant, in ieder geval niet vandaag zal plaatsvinden.

Aan de orde is het **debat** over **het kabinetsstandpunt ten aanzien van het rapport "Dynamiek in islamitisch activisme" van de Wetenschappelijke Raad voor het Regeringsbeleid (WRR) (30800-VI, nr. 115)**.

□

De heer **Wilders** (PVV): Mevrouw de voorzitter. Om te beginnen mijn oprechte dank aan u persoonlijk omdat u op mijn verjaardag vandaag een debat over de islam heeft gepland. Een mooier cadeau had ik mij niet kunnen wensen!

Ongeveer 1400 jaar geleden is ons de oorlog verklaard door een ideologie van haat en geweld die toen ontstond en werd verkondigd door een barbaar die zichzelf profeet Mohammed noemde. Ik heb het over de islam.

Laat ik beginnen met het fundament van de islam: de Koran. De plicht voor alle moslims om strijd te leveren tegen niet-moslims vormt het kernthema van de Koran, door mij eerder de islamitische Mein Kampf genoemd. Met strijd wordt bedoeld: oorlog, jihad. De Koran is vooral een krijgsboek waarin opgeroepen wordt om niet-moslims af te slachten, te braden en om bloedbaden onder hen aan te richten. Joden worden vergeleken met apen en zwijnen. Mensen die geloven dat Jezus Christus de zoon van God is, moeten volgens de Koran bestreden worden (Soera 9, vers 30).

Het Westen kent geen problemen met het joden- en christendom, maar wel met de islam. Moslims kunnen de teksten in de Koran die eeuwig gelden voor alle moslims, ook vandaag de dag nog beschouwen als een "license to kill" en helaas gebeurt dit ook. De inhoud van de Koran is zo geformuleerd dat de bevelen gericht zijn aan moslims van alle tijden, dus ook aan de moslims van nu. In tegenstelling tot teksten in bijvoorbeeld de Bijbel die juist zijn geformuleerd als historische verhalen waarbij de gebeurtenissen in een ver verleden in een context worden geplaatst, is dit bij de Koran niet het geval.

Het waren dan ook moslims en geen joden of christenen die de catastrofale terroristische aanslagen in New York, Madrid en Londen pleegden en niet voor niets werd Theo van Gogh door de moslim Mohammed Bouyeri op beestachtige wijze afgeslacht.

De heer **Dijsselbloem** (PvdA): Mag ik de heer Wilders vragen waarom hij teksten die 1400 jaar oud zijn als volstrekte waarheid en absoluut en voor eeuwig beschouwt? Sluit hij zich daarmee aan bij Osama bin Laden en waarom doet hij dat?

De heer **Wilders** (PVV): Ik doe dat niet, maar heel veel mensen doen dit wel en niet alleen de Osama's bin Laden van deze wereld. Ik weet niet waar de heer Dijsselbloem de afgelopen tien of twintig jaar is geweest, maar als hij om zich heen kijkt in Nederland en buiten Nederland dan zal hij zien dat of het nu gaat over Madrid, Londen, de Filipijnen, Bali of de moord in Amsterdam, er vandaag de dag nog steeds moslims zijn die de teksten van de Koran letterlijk nemen en dat ook moeten doen en daardoor de meest verschrikkelijke dingen doen.

De heer **Dijsselbloem** (PvdA): De heer Wilders hoeft mij niet te vertellen dat er extremisten en idioten zijn die inderdaad met die teksten in de hand tot vreselijke extremiteiten overgaan. De vraag is waarom hij hen gelijk geeft. De vraag is waarom hij hen gelijk geeft in hun interpretatie van die teksten, in de wijze waarop zij menen met die teksten te moeten omgaan. De vraag is waarom hij hen iedere keer weer gelijk geeft.

De heer **Wilders** (PVV): Omdat de Koran een extremistisch boek is en omdat het een extremistisch boek is, moet het worden verboden. De heer Dijsselbloem heeft gelijk: dit betekent natuurlijk niet dat alle moslims extremistisch zijn. Ik maak ook een onderscheid tussen het geloof en de mensen. Ook de Partij voor de Vrijheid heeft geen probleem met moslims die zich houden aan de wet en aan onze Nederlandse rechtsstaat, maar helaas is dit niet het hele verhaal, want de islam, de Koran, de hadith deugen niet en zijn intrinsiek gewelddadig.

De heer **Dijsselbloem** (PvdA): Ik vind het een gewelddige winst dat de heer Wilders een nuance toelaat in zijn betoog. Dat zijn wij al lange tijd niet van hem gewend. Waarom schreef hij echter onlangs, deze zomer, in de krant dat gematigde moslims niet bestaan? Hoe verhoudt dat zich tot zijn tekst nu waarin hij zegt dat hij onderscheid maakt?

De heer **Wilders** (PVV): Ik maak onderscheid tussen de mens en het geloof, maar gematigde moslims bestaan niet. Dat zeg ik niet, maar dat zegt de Koran. Ik zal dadelijk uitleggen waarom de Koran vindt dat er geen gematigde moslims bestaan.

De heer **Dijsselbloem** (PvdA): Heeft de heer Wilders theologie gestudeerd? Is hij islamdeskundige geworden? Waarom begeeft hij zich elke keer weer in een theologisch debat waarin politici naar mijn mening helemaal niet thuis horen? Tenzij het zijn nieuwe vak is. Waarom denkt hij dat hij als het ware "de juiste" interpretatie van de islam te pakken heeft? Is het niet opmerkelijk dat die interpretatie precies gelijk is aan die van de extremisten die hij zegt te bestrijden?

De heer **Wilders** (PVV): Dat is niet opmerkelijk. Het betekent dat de Koran helaas intrinsiek gewelddadig is en niet deugt. Dat is het probleem met het hele boek de

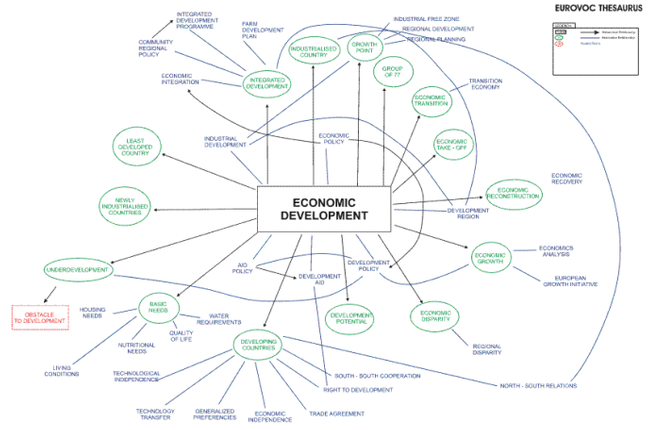


Image from: http://www.iskoi.org/doc/development.htm

Research Questions

- What are the different possible **sources of evidence** for automatic indexing of political text?

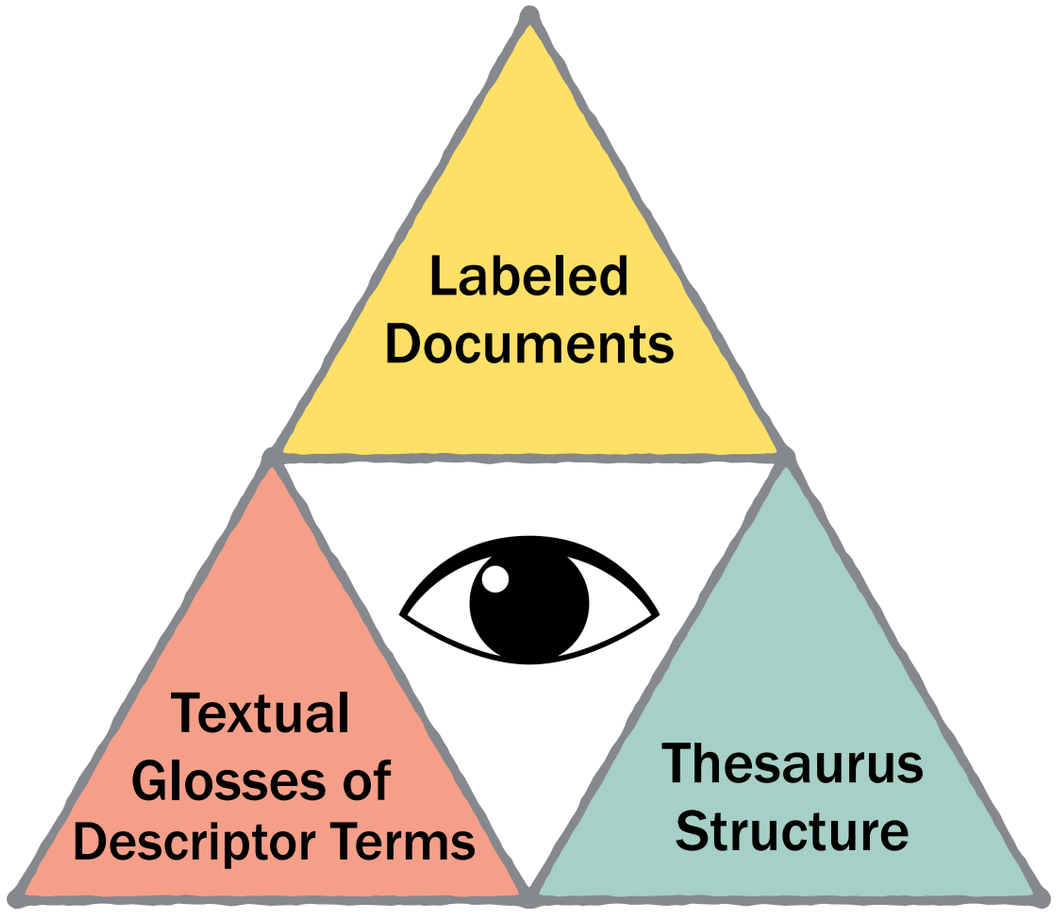
- How effective is a **learning to rank** (LTR) approach integrating a variety of sources of information as features?

- What is the **relative importance** of each of these sources of information for indexing political text?

- Can we build a **lean-and-mean system** that approximate the effectiveness of the large LTR system?

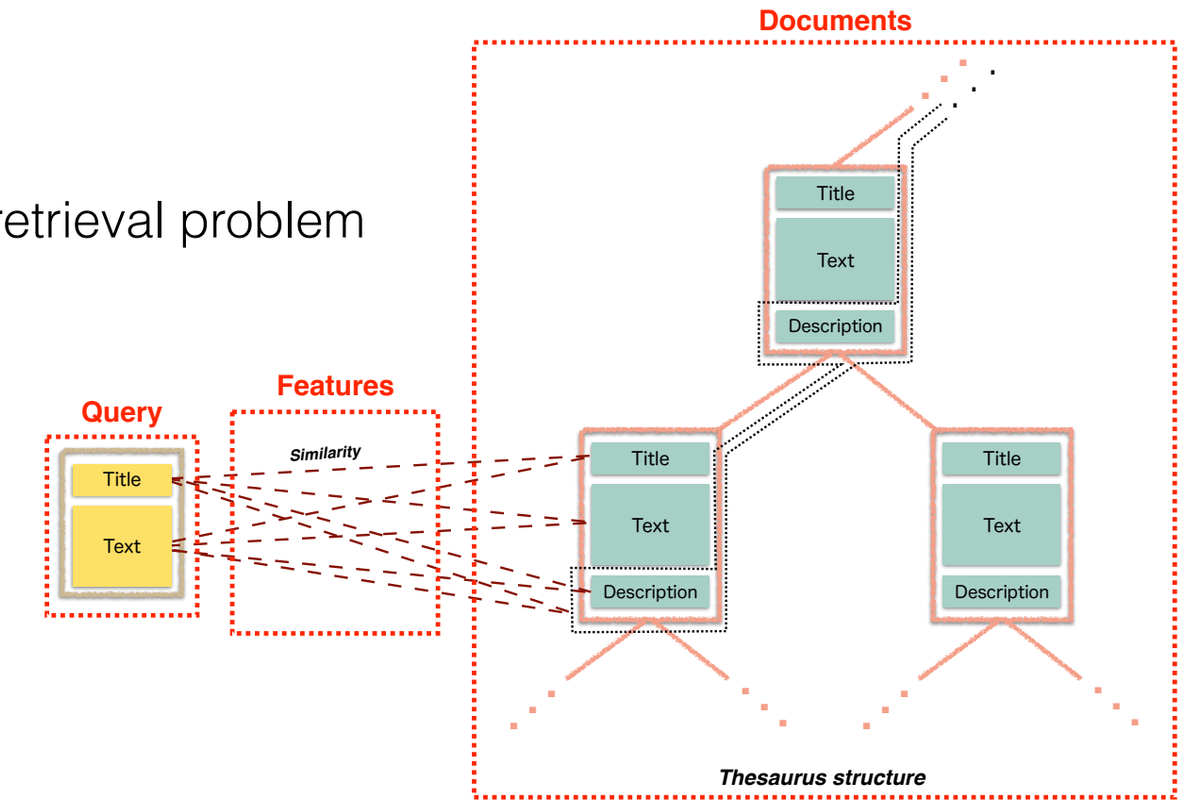
Sources of Evidence

- There are different sources of evidence for the selection of appropriate indexing terms for political documents:



Sources of Evidence Combination

- **Learning to rank**
  - Mapping the problem to an information retrieval problem
    - Documents as *queries*
    - Descriptor terms as *document*
  - Features:
    - Query dependent features (Similarity of query and documents)
    - Query independent features (Prior knowledge)



Features

- **Query dependent features**
  - Model of documents:
    - ***Model_D = <M(title_D), M(text_D)>***
  - Model of descriptor terms:
    - ***Model_DT = <M(title_DT), M(text_DT), M(gloss_DT),M(anc-gloss_DT) >***
  - Eight possible combinations of a document and descriptor term submodel
    - Different IR Similarity Measures (Language modelling similarity based on KL-divergence using Dirichlet smoothing, the same run using Jelinek-Mercer smoothing, Okapi-BM25)
- **Query independent features**
  - Popularity
  - Generality
  - Ambiguity


Rouvoet

hard bezig en ik verwijs dan ook naar het bestuursakkoord met de gemeenten om te bezien of wij een nieuwe balans kunnen vinden in hetgeen lokaal kan gebeuren – preventief, licht en dicht bij huis – en hetgeen echt op provinciaal niveau moet gebeuren, bijvoorbeeld op het gebied van ambulante zorg. Dat geldt ook voor opvoedingsondersteuning. Er gebeurt al heel veel aan opvoedingsondersteuning in het lokale domein. Daar is geen indicatie voor nodig.

De beraadslaging wordt gesloten.

De vergadering wordt enkele ogenblikken geschorst.

Voorzitter: Verbeet

De **voorzitter**: Ik deel u mede dat het debat dat tijdens de regeling van werkzaamheden is aangekondigd door mevrouw Kant, in ieder geval niet vandaag zal plaatsvinden.

Aan de orde is het **debat** over **het kabinetsstandpunt ten aanzien van het rapport "Dynamiek in islamitisch activisme" van de Wetenschappelijke Raad voor het Regeringsbeleid (WRR) (30800-VI, nr. 115)**.

De heer **Wilders** (PVV): Mevrouw de Voorzitter, in het beginnen mijn oprechte dank aan u persoonlijk omdat u op mijn verjaardag vandaag een debat over de islam heeft gepland. Een mooier cadeau had ik mij niet kunnen wensen!

Ongeveer 1400 jaar geleden is ons de oorlog verklaard door een ideologie van haat en geweld die toen ontstond en werd verkondigd door een barbaar die zichzelf profeet Mohammed noemde. Ik heb het over de islam.

Laat ik beginnen met het fundament van de islam: de Koran. De plicht voor alle moslims om strijd te leveren tegen niet-moslims vormt het kernthema van de Koran, door mij eerder de islamitische Mein Kampf genoemd. Met strijd wordt bedoeld: oorlog, jihad. De Koran is vooral een krijgsboek waarin opgeroepen wordt om niet-moslims af te slachten, te braden en om bloedbaden onder hen aan te richten. Joden worden vergeleken met apen en zwijnen. Mensen die geloven dat Jezus Christus de zoon van God is, moeten volgens de Koran bestreden worden (Soera 9, vers 30).

Het Westen kent geen problemen met het joden- en christendom, maar wel met de islam. Moslims kunnen de teksten in de Koran die eeuwig gelden voor alle moslims, ook vandaag de dag nog beschouwen als een "license to kill" en helaas gebeurt dit ook. De inhoud van de Koran is zo geformuleerd dat de bevelen gericht zijn aan moslims van alle tijden, dus ook aan de moslims van nu. In tegenstelling tot teksten in bijvoorbeeld de Bijbel die juist zijn geformuleerd als historische verhalen waarbij de gebeurtenissen in een ver verleden in een context worden geplaatst, is dit bij de Koran niet het geval.

Het waren dan ook moslims en geen joden of christenen die de catastrofale terroristische aanslagen in New York, Madrid en Londen pleegden en niet voor niets werd Theo van Gogh door de moslim Mohammed Bouyeri op beestachtige wijze afgeslacht.

De heer **Dijkhoff** (PvdA): Mag ik de heer Wilders vragen waarom hij teksten die 1400 jaar oud zijn als volstrekte waarheid en absoluut en voor eeuwig beschouwt? Sluit hij zich daarmee aan bij Osama bin Laden en waarom doet hij dat?

De heer **Wilders** (PVV): Ik doe dat niet, maar heel veel mensen doen dit wel en niet alleen de Osama's bin Laden van deze wereld. Ik weet niet waar de heer Dijsselbloem de afgelopen tien of twintig jaar is geweest, maar als hij om zich heen kijkt in Nederland en buiten Nederland dan zal hij zien dat of het nu gaat over Madrid, Londen, de Filippijnen, Bali of de moord in Amsterdam, er vandaag de dag nog steeds moslims zijn die de teksten van de Koran letterlijk nemen en dat ook moeten doen en daardoor de meest verschrikkelijke dingen doen.

De heer **Dijsselbloem** (PvdA): De heer Wilders hoeft mij niet te vertellen dat er extremisten en idioten zijn die inderdaad met die teksten in de hand tot vreselijke extremiteiten overgaan. De vraag is waarom hij hen gelijk geeft. De vraag is waarom hij hen gelijk geeft in hun interpretatie van die teksten, in de wijze waarop zij menen met die teksten te moeten omgaan. De vraag is waarom hij hen iedere keer weer gelijk geeft.

De heer **Wilders** (PVV): Omdat de Koran een extremistisch boek is en omdat het een extremistisch boek is, moet het worden verboden. De heer Dijsselbloem heeft gelijk: dit betekent natuurlijk niet dat alle moslims extremistisch zijn. Ik maak ook een onderscheid tussen het geloof en de mensen. Ook de Partij voor de Vrijheid heeft geen probleem met moslims die zich houden aan de wet en aan onze Nederlandse rechtsstaat, maar helaas is dit niet het hele verhaal, want de islam, de Koran, de hadith deugen niet en zijn intrinsiek gewelddadig.

De heer **Dijsselbloem** (PvdA): Ik vind het een gewelddige winst dat de heer Wilders een nuance toelaat in zijn betoog. Dat zijn wij al lange tijd niet van hem gewend. Waarom schreef hij echter onlangs, deze zomer, in de krant dat gematigde moslims niet bestaan? Hoe verhoudt dat zich tot zijn tekst nu waarin hij zegt dat hij onderscheid maakt?

De heer **Wilders** (PVV): Ik maak onderscheid tussen de mens en het geloof, maar gematigde moslims bestaan niet. Dat zeg ik niet, maar dat zegt de Koran. Ik zal dadelijk uitleggen waarom de Koran vindt dat er geen gematigde moslims bestaan.

De heer **Dijsselbloem** (PvdA): Heeft de heer Wilders theologie gestudeerd? Is hij islamdeskundige geworden? Waarom begeeft hij zich elke keer weer in een theologisch debat waarin politici naar mijn mening helemaal niet thuis horen? Tenzij het zijn nieuwe vak is. Waarom denkt hij dat hij als het ware "de juiste" interpretatie van de islam te pakken heeft? Is het niet opmerkelijk dat die interpretatie precies gelijk is aan die van de extremisten die hij zegt te bestrijden?

De heer **Wilders** (PVV): Dat is niet opmerkelijk. Het betekent dat de Koran helaas intrinsiek gewelddadig is en niet deugt. Dat is het probleem met het hele boek de

Experiments

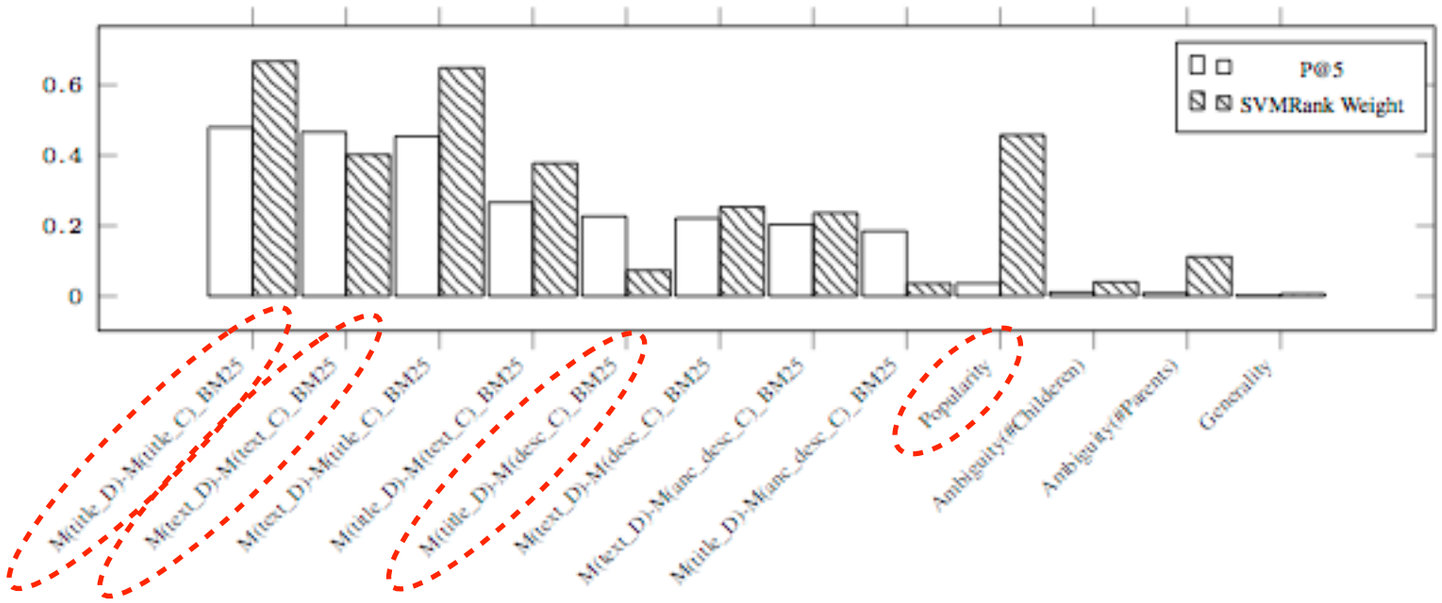
- Data Collection:
  - JRC-Acquis dataset
    - The documents of this corpus have been manually labeled with *EuroVoc* concept

- **Effectiveness of LTR:**
  - Performance of LTR over JEX as baseline system

Method	P@5 (%Diff.)	Recall@5 (%Diff.)
JEX	0.4353	0.4863
BM25-Titles	0.4798 (10%)▲	0.5064 (4%)▲
LTR-all	0.5206 (20%)▲	0.5467 (12%)▲

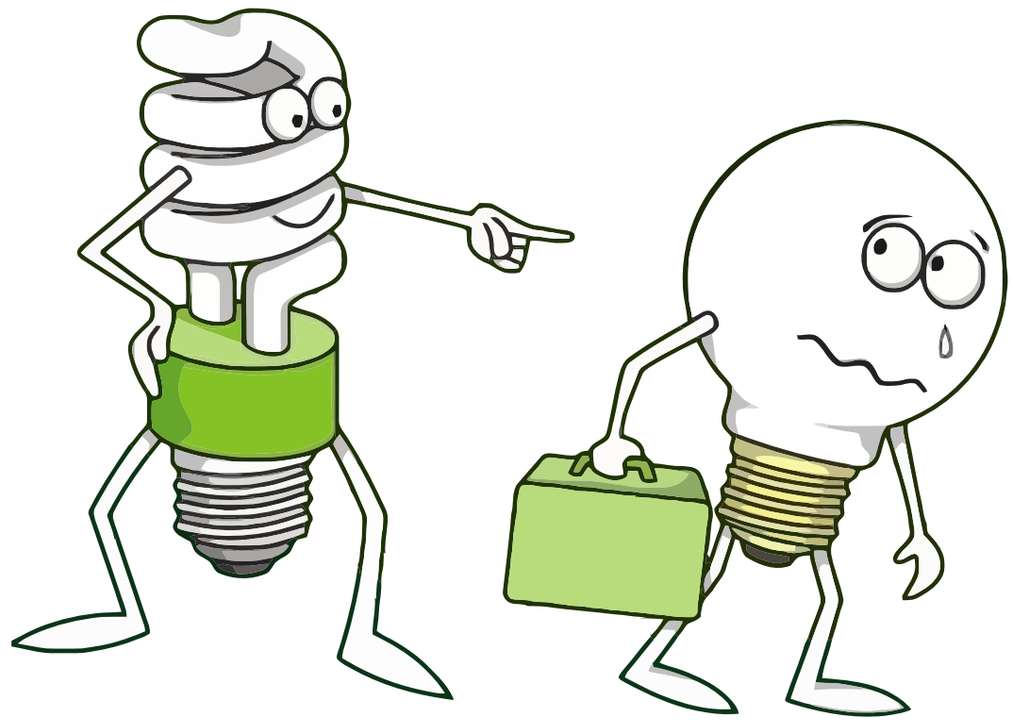
Experiments

- Feature Analysis: **Importance of Different Information Sources**



Experiments

- **Lean-and-Mean** system (Only 4 features):
  - *text submodel of documents with **all text***
  - ***titles** only*
  - *textual **glosses** of descriptor terms*
  - ***popularity** of descriptor terms*



Method	P@5 (%Diff.)	Recall@5 (%Diff.)
LTR-ALL	0.5206 (-)	0.5467 (-)
LTR-TTGP	0.5058 (-3%)	0.5301 (-3%)

Conclusion

- Our main findings:
  - Using a learning to rank (**LTR**) approach **integrating all features** significantly improves the performance of indexing than previous systems.
  - Analysis of feature weights:
    - **relative importance of various sources of evidence**
  - **Lean-and-mean system**:
    - performs at **97%** of the large LTR model

Thank you