Xart system: discovering and extracting correlated arguments of n-ary relations from text

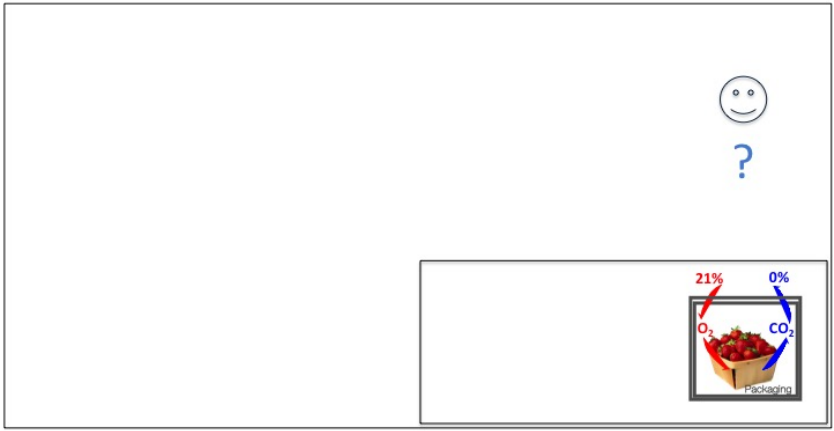
Lilia Berrahou, LIRMM - INRA - Montpellier
Patrice Buche, INRA - LIRMM - UMR IATE - Montpellier
Juliette Dibie, INRA - AgroParisTech - UMR MIA - Paris
Mathieu Roche, CIRAD - LIRMM - UMR TETIS - Montpellier

June 13, 2016
WIMS 2016 - Nîmes, France

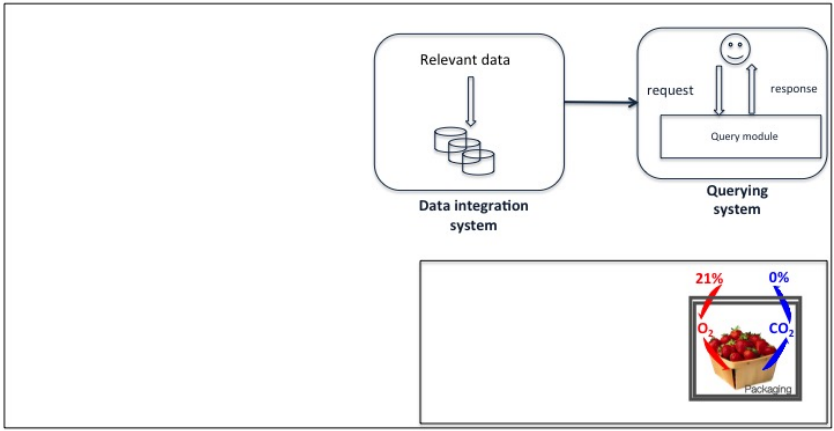
Outline

Information extraction

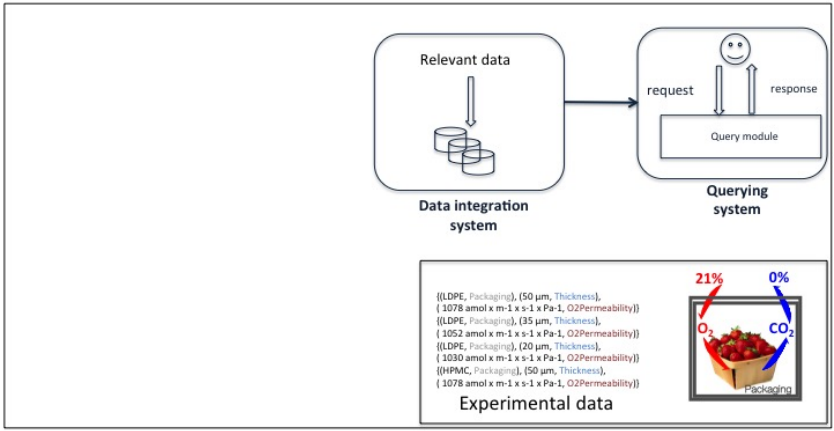
Information extraction



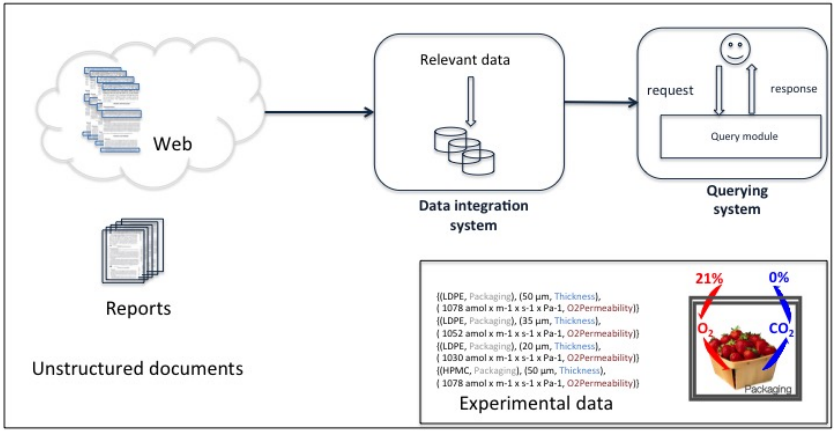
Information extraction



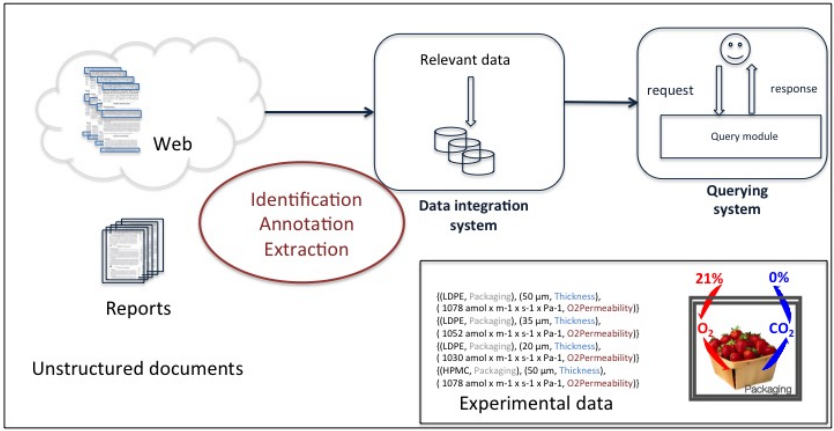
Information extraction



Information extraction



Information extraction



Examples

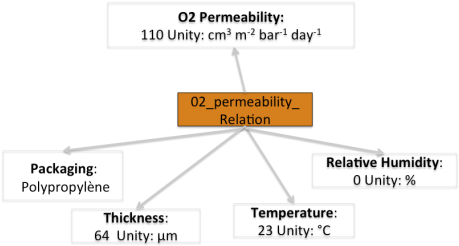
(1) Eight apple wedges were packaged into polypropylene trays and wrap-sealed using a 64 μm thickness polypropylene film with a permeability to oxygen of 110 $cm^3\ m^{-2}\ bar^{-1}\ day^{-1}$ at 23 $^\circ\ C$ and 0 % RH

(2) The A380-800 has a 150 *tons* of transport capacity, a 15 400 *kilometers* of flying range that allow a non-stop New York-Hong Kong flight with a 900 km/h up to 1012 km/h of speed

- **Relevant information:**
  - **Studied objects**: *polypropylene* film, *A380-800* plane
  - **Features**: thickness, permeability, capacity, flying range...
  - **Attributes**: numerical values and units of measure

N-ary relation

- **Relevant information**:
  - A set of quantitative arguments
  - A set of attributes: numerical values and units
  - A studied object: symbolic argument
- Example: Eight apple wedges were packaged into polypropylene trays and wrap-sealed using a 64 μm thickness polypropylene film with a permeability to oxygen of 110 $cm^3\ m^{-2}\ bar^{-1}\ day^{-1}$ at 23 $^\circ$ C and 0 % RH.

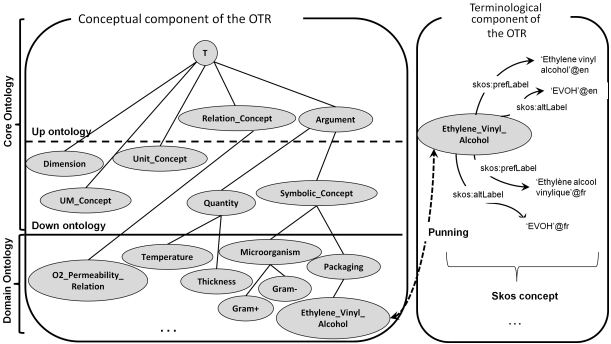


Excerpt of O2_Permeability relation instance

An ontology for n-ary relation representation

two components:

- Conceptual component:
  - Core ontology: generic n-ary relation representation
  - Domain ontology: specific concepts of a given application domain
- Terminological component



Excerpt of the ontology in food packaging domain

Challenges

- Argument expression in several sentences or paragraphs
- Implicit and various forms of expression

Question

Can we find out implicit patterns in the expression of arguments?

N-ary relations

e.g. [Cohen et al., 2009], [Buyko et al., 2009], [Nguyen et al., 2010], [McGrath et al., 2011], [Hawizy et al., 2011], [Ghersedine et al., 2012], [Minard, 2012]

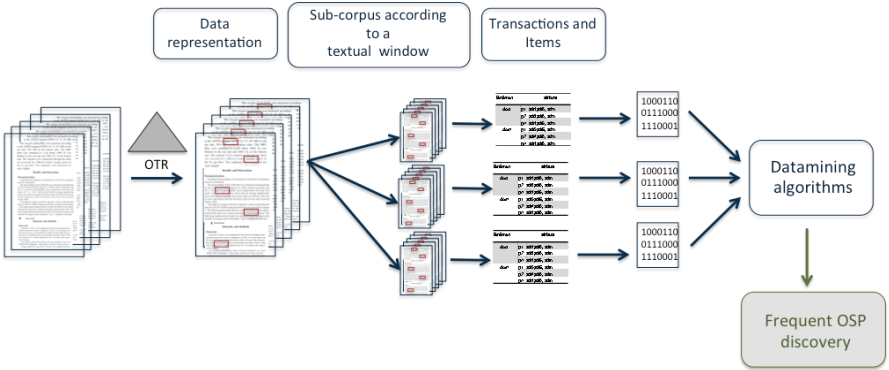
1. **Argument identification**
   - External ressources (lexicon, ontology, thesaurus)
2. **Trigger word identification**
   - Dictionaries
   - Rule-based method
   - Supervised learning methods (classification)
3. **Argument linking**
   - Rule-based method
   - Supervised learning methods (classification)

Background approaches: independant steps and arguments

Our approach: Discovering implicit patterns between arguments

Xart system

KDP

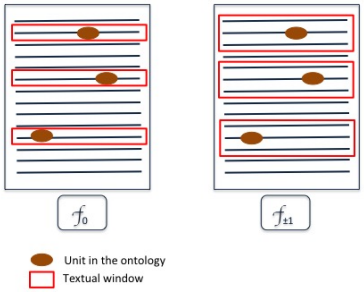

Data representation
Sub-corpus according to a textual window
Transactions and Items
OTR
1000110
0111000
1110001
1000110
0111000
1110001
1000110
0111000
1110001
Datamining algorithms
Frequent OSP discovery

Assumptions

Units of measure in the ontology help to define relevant contexts for discovering arguments [Berrahou et al., 2013]

Relevant textual units:

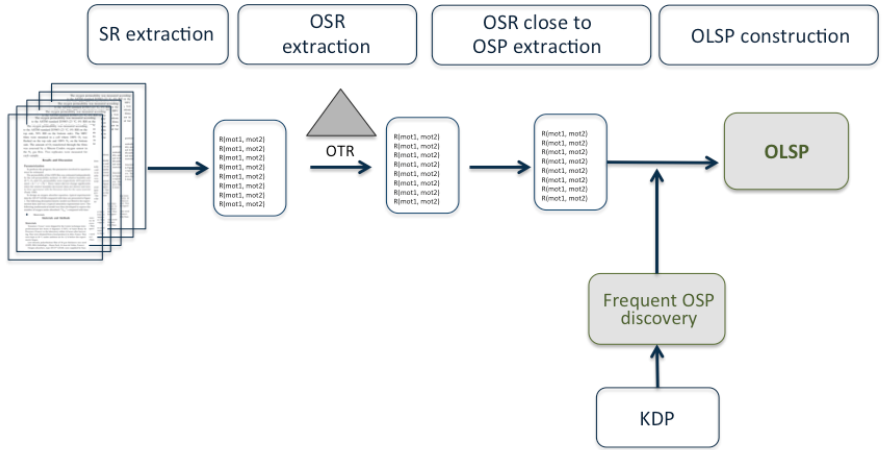
- The pivot sentence
- The textual window



KDP driven by an ontology

- New data representation of arguments and attributes to improve data expressiveness in text
  - Eight apple wedges were packaged into polypropylene < *packaging* > trays and wrap-sealed using a 64 < *numvalthick* > μm < *um* > thickness < *quantity* > polypropylene < *packaging* > film with a permeability to oxygen < *quantity* > of 110 < *numvalperm* > $cm^3\ m^{-2}\ bar^{-1}\ day^{-1}$ < *um* > at 23 < *numvaltemp* > $^\circ$ C < *um* > and 0 < *numvalrh* > % < *um* > RH < *quantity* >.
- Several sub-corpus assessed according to textual windows
- Data mining step to extract Ontological sequential patterns (OSP)
  - Ontological Sequential Pattern (OSP) is a frequent sub-ontological sequence characterized by a support and extracted from a given sub-corpus
  - Example: <(packaging)(numvalthick um)> is an OSP extracted in the sub-corpus $f_{\pm 1}$ with S = 0.5

Hybrid approach



Ontological linguistic sequential pattern construction

1. **SR extraction**: Each sentence of the corpus is parsed
2. **SR close to the ontology (OSR)**: Each SR that contains at least one term denoting a concept
   - Example: NN(thickness, film) is an OSR with NN: noun compound modifier
3. **OSR close to the OSP**: Each OSR that links terms expressing correlated arguments
   - Example: prep_of(thickness, LDPE) is close to the OSP <(packaging)(numvalthick um)> with LDPE a term denoting < packaging > concept
4. **OLSP construction**: Linguistic structures from relevant OSR are integrated in the OSP
   - Example: < (packaging) film thickness (numvalthick um) >
   - Example of identified arguments: *mango films thickness was 0.17 ± 0.02 mm*

Experiments and results

Results

Ontological sequential patterns:

Textual window	Ontological sequential pattern	Support
$f_{\pm 1}$	<**(packaging)(numvalthick um)**> <(packaging)(quantity)(permeability)> <**(packaging)(permeability)**>	0.5 0.5 0.6
f_0	<(pressure)(water permeability)> <(oxygen permeability)(pressure)>	0.05 0.05
$\bigcap f_n$	<**(numvaltemp)(numvalrh%)**> <(packaging)(numvalthick)> <**(packaging)(numvaltemp °c)**>	>0.05

- Correlated arguments discovered from food packaging corpus:
  - Thickness and packaging
  - Temperature and relative humidity
  - Permeability and partial pressure
  - Trigger word of the relation: packaging argument

Argument identification:

Evaluation type	OSP			OLSP		
	Precision	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*
General evaluation	0.5	0.8	**0.6**	0.7	0.8	**0.6**
packaging and *thickness*	0.4	0.9	0.5	0.7	0.9	**0.8**
temperature and *relative humidity*	0.3	0.9	0.4	0.8	0.7	**0.7**
n > 2 correlated arguments	0.6	0.6	**0.6**	0.7	0.6	**0.6**

Conclusion

Xart system is based on a hybrid approach

- Discovering OSP of correlated arguments = patterns of expression in text
- Constructing OLSP = identification of correlated arguments in text

Future work:

- Using Xart system on a new corpus of quantitative data (biorefinery domain)
- Integrating OLSP in an annotation plateform

Thank you for your attention!

References I

Berrahou, S. L., Buche, P., Dibie-Barthélemy, J., and Roche, M. (2013).

How to extract unit of measure in scientific documents?

In *KDIR/KMIS 2013 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval and the International Conference on Knowledge Management and Information Sharing, Vilamoura, Algarve, Portugal, 19 - 22 September, 2013*, pages 249–256.

Buyko, E., Faessler, E., Wermter, J., and Hahn, U. (2009).

Event extraction from trimmed dependency graphs.

In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cohen, K. B., Verspoor, K., Johnson, H. L., Roeder, C., Ogren, P. V., Baumgartner, Jr., W. A., White, E., Tipney, H., and Hunter, L. (2009).

High-precision biological event extraction with a concept recognizer.

In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP '09, pages 50–58, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ghersedine, A., Buche, P., Dibie-Barthélemy, J., Hermandez, N., and Kamel, M. (2012).

Extraction de relations n-aires interphrastiques guidée par une RTO.

In *CORIA*, pages 179–190.

Hawizy, L., Jessop, D., Adams, N., and Murray-Rust, P. (2011).

ChemicalTagger: a tool for semantic text-mining in chemistry.

Journal of cheminformatics, 3(1):17.

References II

McGrath, L. R., Domico, K., Corley, C. D., and Webb-Robertson, B.-J. (2011).
Complex biological event extraction from full text using signatures of linguistic and semantic features.
In *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task '11, pages 130–137, Stroudsburg, PA, USA. Association for Computational Linguistics.

Minard, A. (2012).
Extraction de relations en domaine de spécialité.
PhD thesis, Ecole doctorale informatique de Paris-Sud.

Nguyen, V. T., Gaio, M., and Sallaberry, C. (2010).
Recherche de relations spatio-temporelles : une méthode basée sur l'analyse de corpus textuels.
CoRR, abs/1002.0577.

Yan, X., Han, J., and Afshar, R. (2003).
Clospan: Mining closed sequential patterns in large databases.
In Barbará, D. and Kamath, C., editors, *SDM*. SIAM.