La sémantique d'un récit : état de l'art et perspectives

***Fionn Murtagh*[1,2], *Adam Ganz*[3]**
[1] Science Foundation Ireland, Wilton Place, Dublin 2, Irlande,
[2] Department of Computer Science, Royal Holloway, University of London Egham TW20 0EX, Angleterre
[3] Department of Media Arts, Royal Holloway, University of London Egham TW20 0EX, Angleterre
fmurtagh@acm.org

Résumé: On s'intéresse ici à la sémantique de l'information sous deux aspects : (i) l'ensemble de toutes les relations binaires, (ii) la reconnaissance et le suivi des changements ainsi que des anomalies. Dans le premier cas, on munit l'espace des textes ou des sous textes d'un côté, et l'espace des mots de l'autre côté, d' une métrique euclidienne dans un cadre commun. Dans le deuxième cas, on modélise l'information par une métrique ultramétrique. Une façon d'aboutir à une représentation euclidienne consiste à faire une analyse des correspondances, qui s'applique par exemple à des tableaux en entrée de fréquences. A partir de la représentation euclidienne, on munit l'espace de l'information d'une ultramétrique qui permet de construire une classification hiérarchique. En l'occurrence dans ce travail, on contraindra l'ultramétrique à respecter l'ordre induit par les séquences temporelles. Après une revue de l'existant pour l'analyse de la sémantique des scripts de films, on s'intéresse à la question suivante : serait-il possible d'analyser de façon analogue la sémantique de la littérature de la recherche.

Mots clés: Analyse des correspondances, classification ascendante hiérarchique, classification sous contrainte de contiguïté, analyse textuelle

Abstract We study two aspects of information semantics: (i) the collection of all relationships, (ii) tracking and spotting anomaly and change. The first is implemented by endowing all relevant information spaces with a Euclidean metric in a common projected space. The second is modeled by an induced ultrametric. A very general way to achieve a Euclidean embedding of different information spaces based on cross-tabulation counts (and from other input data formats) is provided by Correspondence Analysis. From there, the induced ultrametric that we are particularly interested in takes a sequential – e.g. temporal – ordering of the data into account. Following a review of approaches adopted in the analysis of filmscript we look at how similar approaches can be applied to the scholarly literature.

Keywords: Correspondence Analysis, hierarchical clustering, contiguity constrained clustering, text analysis

1 Analyse du récit

1.1 Introduction

L'analyse et la fouille des données doit relever principalement les défis suivants :

– Des grandes masses de données, sous forme textuelle et autre, doivent être exploitées comme base décisionnelle. L'analyse des correspondances fournit des réponses pour l'analyse de telles données multidimensionnelles, quantitatives ou nominales ; en un mot, hétérogènes.
– Les structures et les rapports évoluent dans le temps.