

Identification de symboles dans des documents déstructurés

Jacques Péré-Laperne*

*1A3i Arcangues 64200 France

Univ. Bordeaux, ESTIA INSTITUTE OF TECHNOLOGY, F-64210 Bidart, France

j.perelaperne@gmail.com,

<http://www.1a3i.com>

Résumé. Nous décrivons une approche originale pour extraire efficacement les symboles graphiques d'un fichier vectoriel (de type PDF par exemple). Après passage d'un espace d'objets graphiques 2D à une chaîne de codes (1D), l'identification des symboles consiste à rechercher une sous-séquence de codes qui se répète dans le fichier d'entrée. Les travaux de la littérature utilisent l'arbre ou le tableau des suffixes, notre algorithme s'appuie sur le principe du tri par paquets pour identifier les répétitions. La taille et la fréquence sont spécifiées par l'utilisateur.

1 Introduction

L'interprétation d'un fichier vectoriel (PDF, PS, EMF, WFM, IGES, STEP, DXF...), donne une suite ordonnée d'objets graphiques. Dans ce travail nous nous intéressons à l'identification des symboles dans les fichiers vectoriels de type schémas ou plans produits par des logiciels de DAO/CAO (Dessin Assisté par Ordinateur/Conception Assistée par Ordinateur) ou de PAO (Publication Assistée par ordinateur). Ces logiciels dessinent toujours les objets graphiques complexes (symboles) ou simples (rectangles, ellipses) de la même façon, et avec le même ordre chronologique. Partant de ce constat, et des travaux de Péré-Laperne et Couture (2017), ainsi que de ceux de Péré-Laperne (2018) qui précisent comment passer de l'espace des objets graphiques 2D à une chaîne de codes en 1D, sachant que les fonctions de transformation sont insensibles aux translations, aux rotations, ou aux homothéties, l'identification des symboles dans ces fichiers revient à rechercher les sous-chaînes de codes dans ces chaînes de codes issues des fichiers d'entrée. Cette identification des symboles est une des étapes de la méthode (A)KDD (Antropocentric Knowledge Discovery in Database) décrite par Péré-Laperne et Couture (2017) pour la restructuration des documents déstructurés. Cette méthode s'appuie sur les premiers travaux de Fayyad et al. (1996) pour l'extraction des connaissances des données.

Dans la section suivante, nous dressons un bref état de l'art. Dans la section 3, nous faisons le lien entre les répétitions et les paquets de l'algorithme de tri par paquet et nous proposons un algorithme linéaire au niveau du temps et de l'espace d'exécution. Dans la section 4 nous présentons les bénéfices de l'algorithme proposé, avant de conclure dans la dernière section.