Articles recension

胡雨軒*

12 juillet 2014

Abstract

This has been redacted in English, as all abstracts should be ∼ in an effort to keep the general meaning of that document intelligible.

This is a recension meant to cover each article in the `ref` directory. First abstract is pasted then I briefly explain why I've chosen to put it in my bibliography. This is aimed at not getting drown in all those deep, difficult yet enthralling articles.

I'll try for each of them to reply to several questions: what can I use from those for my own aims? What are new suggested projects? Does it make use of ideas I ought use for my own?

Table des matières

* <p2b.fac@gmail.com>

Introduction

1 "Efficient learning strategy of Chinese characters based on network approach"

1.1 Résumé

Based on network analysis of hierarchical structural relations among Chinese characters, we develop an efficient learning strategy of Chinese characters. We regard a more efficient learning method if one learns the same number of useful Chinese characters in less effort or time. We construct a node-weighted network of Chinese characters, where character usage frequencies are used as node weights. Using this hierarchical node-weighted network, we propose a new learning method, the distributed node weight (DNW) strategy, which is based on a new measure of nodes' importance that takes into account both the weight of the nodes and the hierarchical structure of the network. Chinese character learning strategies, particularly their learning order, are analyzed as dynamical processes over the network. We compare the efficiency of three theoretical learning methods and two commonly used methods from mainstream Chinese textbooks, one for Chinese elementary school students and the other for students learning Chinese as a second language. We find that the DNW method significantly outperforms the others, implying that the efficiency of current learning methods of major textbooks can be greatly improved.

1.2 Réaction

C'est le premier article que j'ai lu. A Taïwan j'avais déjà en tête de construire un graphe (ils appellent ça un réseau), et de trier les caractères (vus comme des nœuds) par degrés. Cet article ajoute une idée intéressante : panacher avec des fréquences.

En voyant leur full map of Chinese characters network j'ai eu envie de l'explorer davantage. Ils disent montrer un minimal spanning tree mais il en existe plusieurs possibles.

Cet article est adossé à http://learnm.org/ qui propose beaucoup de matériau.

Ils n'ont pas donné de tables d'adjacence sur leur site mais plutôt des listes d'adjacence. Pas fou ! Ca m'inquiétait un peu de voir un tableau contenant à peu près 3500^2 zéros !

[Maths avec les mains] La figure 5 présente deux cas : d'abord une courbe qui

finit droite puis une courbe qui penche (car on intègre la fréquence). Soit b la valeur maximum atteinte par cette courbe et x la quantité de fréquence qu'on intègre. Exprimer $b(x)$. A relire

1.3 Pistes d'application

Peut-on montrer que le choix des clefs par les anciens est un optimal ? de quel type et selon quels critères ?

Accès aux caractères On peut aussi regarder combien de caractères il faut pour accéder à tous les caractères. Par exemple avec trois caractères je peux apprendre toutes leur combinaisons possibles mais pas plus [1]. Quelle est l'évolution de la taille du jeu de caractères en fonction du nombre de caractères auxquels on veut accéder ? La tête de cette fonction doit être intéressante. La taille du jeu n'est pas forcément très impressionnante : les lettrés ont proposé 214 pour Kangxi, il y en a souvent moins pour les dictionnaires condensés modernes.

DNW Je pense que leur idée de prendre en compte both the weight of the nodes and the hierarchical structure of the network est bonne. Cependant un étudiant étudie (sic) et passe des examens. Un examen connu dans le monde sinophone est le HSK. Je ne connais pas l'équivalent taïwanais mais je suppose que le HSK a une déclinaison en caractères non simplifiés. Le site http://hskhsk.com offre alors une démarche intéressante. Il faudrait s'en inspirer et faire des listes progressives qui pour chaque niveau contiennent tous les caractères requis plus un minimum. Ce minimum serait tous les caractères intermédiaires nécessaires, les clefs et les caractères proches.

Qu'est-ce qu'un caractère proche ? Il faudrait voir dans la structure du chinois. Par exemple, un caractère ayant les mêmes composants mais pas la même structure peut être considérer proche. A cet effet il doit être instructif se renseigner sur les types de décomposition.

Erreurs "Visually and phonologically similar characters in incorrect Chinese words: Analyses, identification, and applications" [2] permet d'apprendre des erreurs standard des étudiants chinois. On pourrait également choisir de proposer en même temps qu'un caractère différents autres pour éviter des erreurs, ou au contraire éloigner le plus possible leur apprentissage pour éviter toute confusion [2] suivant le type d'erreurs..

1. Le nombre est élevé mais il faut restreindre à l'ensemble de caractères existants.
2. Exemples : droite et gauche (apprendre ensemble ou séparément ?) ; aimer et détester

Base de données Dans le paragraphe de l'accès au caractères, une application standard peut être d'étudier et d'optimiser the Academia Sinica's ???????? Chinese character structure database. Ca serait vraiment un travail sur la base de données pour trouver le meilleur type. Base de données en graphe ? autres ?

Conclusion Cet article a le bon goût de susciter plein de questions. Loin de n'être pas intéressant, l'étude des caractères chinois est parfaitement possible en restant en informatique et il semble même que l'informatique soit la manière reine d'étudier les caractères chinois en revenant à sa définition fondamentale : science de l'information.

Théorie des graphes, bases de données, combinatoire et même en cherchant un peu théorie de l'apprentissage et intelligence artificielle : la langue chinoise est passionnante et permet manifestement à tout un chacun de s'éclater.

Note personnelle Ca me rappelle ce que me disait un animateur[3] de Mathématic Park : à partir d'un certain moment quand on s'intéresse à un domaine on prend des livres et on apprend tout seul ; il avait manifestement raison. Je redoute d'avoir à me plonger dans des mathématiques de MP ! mais dans le même temps je suis curieux de voir ce qui pourrait m'y emmener. Si le travail de chercheur consiste à apprendre tout ce qui peut le mener à ses fins alors il n'y a pas de métier facile mais celui-ci en est un beau.

Conclusion

About LaTeX composition and misc

Rajouter une fonction qui affiche une étoile dans la marge pour indiquer les points à relire et donner une liste de ces étoiles.

C'est bizarre, les sinogrammes ne sont pas reconnus quand ils sont dans `foreignlanguage{}` mais le sont bien à l'extérieur.

Il faut régler ce p*** de problème avec les chemins. Il faut avoir une fonction `\root` qui donnerait à la compilation le chemin absolu vers la racine du dépôt github. Déclarer des répertoires et organiser les préambules et autres serait bien plus facile ! On pourrait saucissonner les préambules pour des parties dans certains dossiers et d'autres ailleurs et ça serait super pour factoriser.

3. þ : celui qui avait une tête sympa au petit nez pointu, que j'avais revu à une lecture inaugurale au collège de France mais également dans un train. Je ne me rappelle pas de son prénom :-(.

Bibliographie

[1] Xiaoyong YAN et al. “Efficient learning strategy of Chinese characters based on network approach”. In : *PloS one* 8.8 (2013), e69745.

Références à traiter

A télécharger [4]

[3] Jiajia HU et Ning WANG. “Graph model of Old Chinese phonological system and computing”. In : *Literary and linguistic computing* (2012), fqs001.

[4] Jiajia HU et Ning WANG. “Complex network perspective on graphic form system of Hanzi”. In : *Literary and linguistic computing* (2013), fqt057.

[5] Shuigeng ZHOU et al. “An empirical study of Chinese language networks”. In : *Physica A : Statistical Mechanics and its Applications* 387.12 (2008), p. 3039–3047.

[6] Jianyu LI et Jie ZHOU. “Chinese character structure analysis based on complex networks”. In : *Physica A : Statistical Mechanics and its Applications* 380 (2007), p. 629–638.

[7] Chad HANSEN. “Chinese language, Chinese philosophy, and “truth” ”. In : *The Journal of Asian Studies* 44.03 (1985), p. 491–519.

[8] Wei LIANG, Yuming SHI et Qiuling HUANG. “Modeling the Chinese language as an evolving network”. In : *Physica A : Statistical Mechanics and its Applications* 393 (2014), p. 268–276.

[9] Shuiyuan YU, Haitao LIU et Chunshan XU. “Statistical properties of Chinese phonemic networks”. In : *Physica A : Statistical Mechanics and its Applications* 390.7 (2011), p. 1370–1380.

[10] Michael E BALES et Stephen B JOHNSON. “Graph theoretic modeling of large-scale semantic networks”. In : *Journal of biomedical informatics* 39.4 (2006), p. 451–464.

[11] Jianyu LI et al. “Chinese lexical networks : The structure, function and formation”. In : *Physica A : Statistical Mechanics and its Applications* 391.21 (2012), p. 5254–5263.

4. Articles et documents auxquels je n’ai pas pu accéder.

[12] Helen H Shen et Chuanren Ke. “Radical awareness and word acquisition among nonnative learners of Chinese”. In : *The Modern Language Journal* 91.1 (2007), p. 97–111.

[13] Biyin Zhang et Danling Peng. “Decomposed storage in the Chinese lexicon”. In : *Advances in psychology* 90 (1992), p. 131–149.

[27] BAI Yi YI JunKai. “Research and test on code-based rare Chinese character input method”. In : *Journal of Beijing University of Chemical Technology (Natural Science Edition)* (2007), S1.

A lire [5]

[2] C-L Liu et al. “Visually and phonologically similar characters in incorrect Chinese words : Analyses, identification, and applications”. In : *ACM Transactions on Asian Language Information Processing (TALIP)* 10.2 (2011), p. 10.

[14] Taran Grant et Arnold G. Kluge. “Transformation Series as an Ideographic Character Concept”. In : *Cladistics* 20.1 (2004), p. 23–31. issn : 1096-0031. doi : 10.1111/j.1096-0031.2004.00003.x. url : http://dx.doi.org/10.1111/j.1096-0031.2004.00003.x.

[15] Jianwei Wang, Lili Rong et Tao Jin. “An empirical study of Chinese word-word language directed network”. In : *Service Operations and Logistics, and Informatics, 2008. IEEE/SOLI 2008. IEEE International Conference on*. T. 1. IEEE. 2008, p. 498–501.

[16] Shixiao Wu et Shijue Zheng. “A Structure Character Modeling for Chinese Character Glyph Description”. In : *Electronic Computer Technology, 2009 International Conference on*. IEEE. 2009, p. 245–248.

[17] Yun Li et Mei Xie. “Chinese character recognition based on character reconstruction”. In : *Communications, Circuits and Systems, 2009. ICCCAS 2009. International Conference on*. IEEE. 2009, p. 460–463.

[18] You-Yang Yu et al. “Chinese language processing with complex network theory”. In : *Computer Science and Software Engineering, 2008 International Conference on*. T. 1. IEEE. 2008, p. 710–713.

5. En d’autres termes : taf

[19] Jingning Ji, Liangrui Peng et Bohan Li. "Graph Model Optimization Based Historical Chinese Character Segmentation Method". In : *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*. IEEE. 2014, p. 282–286.

[20] WB Deng et al. "Rank-frequency relation for Chinese characters". In : *arXiv preprint arXiv :1309.1536* (2013).

[21] Derming Juang et al. "Resolving the unencoded character problem for Chinese digital libraries". In : *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*. IEEE. 2005, p. 311–319.

[22] Candy LK Yiu et Wai Wong. "Chinese character synthesis using METAPOST". In : *In proceedings of TUG*. 2003, p. 85–93.

[23] Hiromichi Fujisawa, Yasuaki Nakano et Kiyomichi Kurino. "Segmentation methods for character recognition : from segmentation to document structure analysis". In : *Proceedings of the IEEE* 80.7 (1992), p. 1079–1092.

[24] Bowen Yu et al. "Statistical Structure Modeling and Optimal Combined Strategy Based Chinese Components Recognition". In : *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*. IEEE. 2012, p. 238–245.

[25] Chen-Yu Lai et al. "A composite approach to handle missing characters on Web interface". In : *ICDAT2004* (2004).

[26] Min Lin, Rou Song et Shi-Li Ge. "A Research on the Stroke-Segment-Mesh (SSM) Glyph Depiction Method of Chinese Character". In : *Advanced Language Processing and Web Information Technology, 2008. ALPIT'08. International Conference on*. IEEE. 2008, p. 269–278.

[28] Matthew Skala. "A Structural Query System for Han Characters". In : *arXiv preprint arXiv :1404.5585* (2014).

[29] EI Le Quan Ha, Ji Ming et FJ Smith. "Extension of Zipf's law to word and character n-grams for English and Chinese". In : *Journal of Computational Linguistics and Chinese Language Processing*. Citeseer. 2003.

[30] Alessandro Giacalone, Martin C Rinard et Thomas W Doeppner Jr. "IDEOSY : An ideographic and interactive program description system". In : *ACM SIGPLAN Notices*. T. 19. 5. ACM. 1984, p. 15–20.

[31] Richard S Cook. "UniHan Variation : Issues and Solutions". In : *23 th Internationalization and Unicode Conference, Prague, Czech Republic*. 2003.

[32] Yannis HARALAMBOUS. "New perspectives in sinographic language processing through the use of character structure". In : *Computational Linguistics and Intelligent Text Processing*. Springer, 2013, p. 201–217.