Articles recension

胡雨軒*

17 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>

References

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

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

Not processed yet

To be downloaded [1]

[5] Jiajia Hu and Ning Wang. "Graph model of Old Chinese phonological system and computing". In: *Literary and linguistic computing* (2012), fqs001.

[6] Jiajia Hu and Ning Wang. "Complex network perspective on graphic form system of Hanzi". In: *Literary and linguistic computing* (2013), fqt057.

[7] Shuigeng Zhou et al. "An empirical study of Chinese language networks". In: *Physica A: Statistical Mechanics and its Applications* 387.12 (2008), pp. 3039–3047.

[8] Jianyu Li and Jie Zhou. "Chinese character structure analysis based on complex networks". In: *Physica A: Statistical Mechanics and its Applications* 380 (2007), pp. 629–638.

[9] Chad Hansen. "Chinese language, Chinese philosophy, and "truth"". In: *The Journal of Asian Studies* 44.03 (1985), pp. 491–519.

[10] Wei Liang, Yuming Shi, and Qiuling Huang. "Modeling the Chinese language as an evolving network". In: *Physica A: Statistical Mechanics and its Applications* 393 (2014), pp. 268–276.

1. Articles and documents I have not successfully got yet. If you have been granted access to some of them, please send me them ;-)

[11] Shuiyuan Yu, Haitao Liu, and Chunshan Xu. "Statistical properties of Chinese phonemic networks". In: *Physica A: Statistical Mechanics and its Applications* 390.7 (2011), pp. 1370–1380.

[12] Michael E Bales and Stephen B Johnson. "Graph theoretic modeling of large-scale semantic networks". In: *Journal of biomedical informatics* 39.4 (2006), pp. 451–464.

[13] Jianyu Li et al. "Chinese lexical networks: The structure, function and formation". In: *Physica A: Statistical Mechanics and its Applications* 391.21 (2012), pp. 5254–5263.

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

[15] Biyin Zhang and Danling Peng. "Decomposed storage in the Chinese lexicon". In: *Advances in psychology* 90 (1992), pp. 131–149.

[28] 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.

To be read [2]

[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.

[4] Taran Grant and Arnold G. Kluge. "Transformation Series as an Ideographic Character Concept". In: *Cladistics* 20.1 (2004), pp. 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.

[16] Jianwei Wang, Lili Rong, and 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*. Vol. 1. IEEE. 2008, pp. 498–501.

[17] Shixiao Wu and Shijue Zheng. "A Structure Character Modeling for Chinese Character Glyph Description". In: *Electronic Computer Technology, 2009 International Conference on*. IEEE. 2009, pp. 245–248.

2. In other way: go to work

[18] Yun Li and Mei Xie. "Chinese character recognition based on character reconstruction". In: *Communications, Circuits and Systems, 2009. ICCCAS 2009. International Conference on*. IEEE. 2009, pp. 460–463.

[19] You-Yang Yu et al. "Chinese language processing with complex network theory". In: *Computer Science and Software Engineering, 2008 International Conference on*. Vol. 1. IEEE. 2008, pp. 710–713.

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

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

[22] 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, pp. 311–319.

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

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

[25] 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, pp. 238–245.

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

[27] Min Lin, Rou Song, and 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, pp. 269–278.

[29] EI Le Quan Ha, Ji Ming, and 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, and Thomas W Doeppner Jr. "IDEOSY: An ideographic and interactive program description system". In: *ACM SIGPLAN Notices*. Vol. 19. 5. ACM. 1984, pp. 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, pp. 201–217.

Introduction

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

1.1 Abstract

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[3]. 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.

3. Le nombre est élevé mais il faut restreindre à l'ensemble de caractères existants.

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 [4] suivant le type d'erreurs..

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 [5] 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.

4. Exemples : droite et gauche (apprendre ensemble ou séparément ?) ; aimer et détester

5. Þ : celui qui avait un 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 :-(.

2 “A Structural Query System for Han Characters”

2.1 Note de version

L’article que je lis est une pré-impression : `preprint arXiv:1404.5585`, peut-être y aura-t-il une nouvelle version ; à surveiller donc. Mais cette version a été publiée le 22 avril 2014 donc c’est récent.

2.2 Abstract

The IDSgrep structural query system for Han character dictionaries is presented. This system includes a data model and syntax for describing the spatial structure of Han characters using Extended Ideographic Description Sequences (EIDSes) based on the Unicode IDS syntax; a language for querying EIDS databases, designed to suit the needs of font developers and foreign language learners; a bit vector index inspired by Bloom filters for faster query operations; a freely available implementation; and format translation from popular third-party IDS and XML character databases. Experimental results are included, with a comparison to other software used for similar applications.

2.3 Ressources en ligne du projet

http://tsukurimashou.sourceforge.jp/
http://tsukurimashou.sourceforge.jp/idsgrep.php.en
http://ansuz.sooke.bc.ca/ Le site de l’auteur. J’y étais déjà allé faire un tour ! D’un certain côté c’est rassurant de retomber sur des choses connues mais ça prouve que je n’intègre pas assez profondément ce que je trouve, snif.

2.4 Réaction

J’ai trouvé cet article assez tard mais son titre m’a appâté et je l’ai lu en second. Outre que cela révèle que je n’ai pas lu d’autres articles que le premier avant de trouver celui-ci ∼∼, j’imaginais pouvoir faire une base de données d’idéogrammes et lui envoyer des requêtes mais l’idée d’élaborer un langage de requête propre et d’améliorer les IDS d’unicode va plus loin, c’est mieux.

Le second lien indique : Tsukurimashou, KanjiVG, CHISE, and EDICT2. cela me fait toujours quelques bases de données en plus. Le corps de l’article (table I) indique des jeux de données.

Eh beh voilà un bel état de l'art ! Le paragraphe 1.1 recense plein de projets dont j'ai déjà eu connaissance et leur apporte un éclairage nouveau. Du coup cela montre que cet article est vraiment bien dans mon cœur de cible. Il parle de WWWJDIC qui utilise un autre jeu de clef que les canoniques. On retrouve l'idée de mon premier brouillon qui décomposait les caractères de deux manières.

Je n'ai pas l'impression qu'il utilise des bases de données graphes mais plutôt un stockage en arbre dans des fichiers textes. Comme indiqué par le nom de son outil, le but d'IDSgrep est de parcourir du texte.

En tout cas c'est cool, ça montre bien qu'il y a de vrais possibilités de faire de l'informatique avec les idéogrammes. Je suis content d'avoir choisi le parcours ingénieur logiciel : l'argument utilisé à l'époque semble toujours aussi bon.

On retrouve dans cet article le vocabulaire d'arbre que j'utilisais dans mon premier brouillon. "A Structural Query System for Han Characters" utilise plutôt le point de vue du réseau mais les deux sont sans doute complémentaires et intéressants.

Je me souviens que j'ai commencé de lire un jour le cours de linguistique générale de SAUSSURE. Ce domaine étant à l'interface entre informatique et linguistique, peut-être ferai-je bien de le relire !

http://arxiv.org/abs/1407.3751?context=cs.IR

Cet article est vraiment technique. Il présente quelques nouvelles idées et surtout les détails techniques de leurs implémentations. Je pense qu'il faudra que je le relise quand je travaillerai vraiment le sujet et que ma tâche principale de lire des articles sera terminée.

Dans le doute. A lire.

2.5 Pistes d'application

Tsukurimashou font project may be a elegant way to generate output for components yet outside of unicode.

Existe-t-il un ensemble de clefs parfait ? comment définir cette perfection ? comment situer l'ensemble de clefs canoniques par rapport à cette définition ? cela ne dépendrait-il pas de l'utilisation ?

Je pense qu'on peut voir comment implémenter ça avec une base de données RDF ou en graphe. Peut-être pourrais-je demander conseil à madame CHIKY ? http ://nick-patch.net/code/slides/ à voir, peut-être l'inclure dans les références en ligne.

3 “Transformation Series as an Ideographic Character Concept”

Son titre m’a attiré. Il pourrait n’être pas tout à fait dans le sujet mais au moins je serai fixé.

Conclusion

Dresser un état synthétique de l’art. Qu’existe-t-il déjà ? que pourrait-il rester à faire ?

4 LaTeX todos

Put websites in references and display it in two-part bibliography splitted with papers and online resources.

Methinks there is some issue with inline ideograms…

Include personal draft in entries.bib under the personal draft section to make it easy to evoke previous steps.