INTER-FACULTY MASTER PROGRAM on
COMPLEX SYSTEMS and NETWORKS
SCHOOL of MATHEMATICS
SCHOOL of BIOLOGY
SCHOOL of GEOLOGY
SCHOOL of ECONOMICS
ARISTOTLE UNIVERSITY of THESSALONIKI

Master Thesis

Title:

Multimedia Representation Using Graph-Based Models and Applications

Αναπαράσταση Πολυμέσων με Μοντέλα Γραφημάτων και Εφαρμογές

Elissavet Batziou

SUPERVISOR: Ioannis Antoniou, Full Professor, Aristotle University of Thessaloniki

CO-SUPERVISOR: Ilias Gialampoukidis, Postdoctoral Researcher, Centre for Research and Technology – Hellas (CERTH)

Thessaloniki, June 2017

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ στα

ΠΟΛΥΠΛΟΚΑ ΣΥΣΤΗΜΑΤΑ και ΔΙΚΤΥΑ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ
ΤΜΗΜΑ ΓΕΩΛΟΓΙΑΣ
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τίτλος Εργασίας:

Αναπαράσταση Πολυμέσων με Μοντέλα Γραφημάτων και Εφαρμογές

Multimedia Representation Using Graph-Based Models and Applications

Ελισσάβετ Μπάτζιου

ΕΠΙΒΛΕΠΩΝ: Ιωάννης Αντωνίου, Καθηγητής, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
ΣΥΝΕΠΙΒΛΕΠΩΝ: Ηλίας Γιαλαμπουκίδης, Μεταδιδακτορικός Ερευνητής, Εθνικό Κέντρο Έρευνας και Τεχνολογικής Ανάπτυξης (ΕΚΕΤΑ)

Εγκρίθηκε από την Τριμελή Εξεταστική Επιτροπή την ____ Ιουνίου 2017.

……	……	……
Ι. Αντωνίου	Η. Γιαλαμπουκίδης	Σ. Βροχίδης
Καθηγητής Α.Π.Θ.	Ερευνητής ΕΚΕΤΑ	Ερευνητής ΕΚΕΤΑ

Θεσσαλονίκη, Ιούνιος 2017

……..
Ελισσάβετ Ι. Μπάτζιου
Πτυχιούχος Μαθηματικός
Πανεπιστήμιο Αιγαίου

ABSTRACT

In this work, we compare different methods for keyword extraction and text clustering using the Bag of Words (BoW) and the Graph of Words (GoW) models, along with their extension in image representation. For that purpose we provide the necessary background from the graph theory and community detection approaches. Firstly, we introduce the basic concepts from graph theory, such as centrality measures and community detection approaches, which are used to represent a multimedia item (text or image) into a vector or a graph representation. Moreover, we discuss the GoW model and how text is represented as a graph. Furthermore, we introduce the construction of a visual vocabulary using a graph of visual words, in analogy to the graph of words in text modeling. We apply these models in text/image collections, in order to examine which method is more effective in real data. We evaluate the BoW and the GoW models in keyword extraction from text. Then, we compare popular clustering methods in public datasets with news articles. Moreover, we compare the proposed Graph of Visual Words (GoVW) model with the Bag of Visual Words (BoVW) model in image collections, where we observe that NMI score increases by 6.68% and 16.11% in both of datasets, using GoVW model. Finally, we demonstrate a qualitative comparison in results of images clustering in order to visualize the results.

KEY WORDS

Graph theory, Text, Image, keyword extraction, community detection, density-based clustering, Image clustering, Image retrieval, Bag of Words, Graph of Words, Bag of Visual Words, Graph of Visual Words

ΠΕΡΙΛΗΨΗ

Στα πλαίσια της παρούσης εργασίας, συγκρίνουμε διάφορες μεθόδους συσταδοποίησης και εξόρυξης λέξεων-κλειδιών από κείμενα χρησιμοποιώντας τα μοντέλα Bag of Words (BoW) και Graph of Words (GoW), αλλά και τις γενικεύσεις τους στην αναπαράσταση εικόνων. Για το σκοπό αυτό παρέχουμε το απαραίτητο υλικό από τη θεωρία γράφων. Αφού εισάγουμε τις βασικές έννοιες από τη θεωρία γραφημάτων, τα μέτρα κεντρικότητας και τις μεθόδους ανίχνευσης κοινοτήτων σε δίκτυο που θα χρησιμοποιήσουμε έτσι ώστε να εξετάσουμε την αποτελεσματικότητα τους, εξηγούμε πως ένα κείμενο μετατρέπεται σε διάνυσμα μέσω του μοντέλου BoW και πώς αυτό αξιοποιείται. Έπειτα αναλύουμε το μοντέλο GoW και εξηγούμε μέσω ποιας διαδικασίας ένα κείμενο μπορεί να μετατραπεί και να γίνει αναπαράσταση αυτού σε γράφημα. Επίσης, κατασκευάζουμε τα αντίστοιχα μοντέλα αναπαράστασης εικόνων με στατιστικές μεθόδους (BoVW) αλλά και με τη θεωρία γραφημάτων (GoVW), και τα αξιολογούμε μέσω πειραμάτων σε πραγματικά δεδομένα. Εφαρμόζουμε όλα αυτά τα μοντέλα και τα μέτρα σε πραγματικές συλλογές κειμένων και εικόνων ώστε να εξετάσουμε ποια μέθοδος είναι περισσότερο αποτελεσματική σε πραγματικά δεδομένα. Πρώτον, αξιολογούμε τις μεθόδους BoW και GoW στην εξόρυξη λέξεων κλειδιών από ένα κείμενο. Δεύτερον, συγκρίνουμε γνωστές μεθόδους συσταδοποίησης κειμένου σε δημόσια διαθέσιμες συλλογές άρθρων ειδήσεων. Τρίτον, συγκρίνουμε τα μοντέλα Graph-of-Visual-Words (GoVW) και Bag-of-Visual-Words (BoVW) σε συλλογές εικόνων που είναι δημόσια διαθέσιμες και παρατηρούμε ότι ο δείκτης NMI αυξάνεται κατά 6.68% και 16.11% στις δυο συλλογές αντίστοιχα χρησιμοποιώντας τη μέθοδο GoVW που προτείνουμε έναντι της BoVW. Ολοκληρώνουμε την σύγκριση με μία ποιοτική σύγκριση των αποτελεσμάτων συσταδοποίησης εικόνων για οπτικοποίηση των αποτελεσμάτων.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

Αναπαράσταση κειμένου και εικόνας σε γράφημα, Ανίχνευση κοινοτήτων, Μέτρα κεντρικότητας δικτύου, Εξόρυξη λέξεων κλειδιών, Συσταδοποίηση κειμένων και εικόνων, μοντέλα GoVW, BoVW, BoW και GoW.

Table of Contents

ΣΥΝΟΨΗ

Στα πλαίσια της παρούσης εργασίας, συγκρίνουμε διάφορες μεθόδους εξόρυξης λέξεων από κείμενα χρησιμοποιώντας τα μοντέλα Bag of Words (BoW) και Graph of Words (GoW). Για το σκοπό αυτό παρέχουμε το απαραίτητο υλικό από τη θεωρία γράφων και το διαχωρισμό κοινοτήτων. Αφού εισάγουμε τις βασικές έννοιες από τη θεωρία γραφημάτων, τα μέτρα κεντρικότητας και τις μεθόδους ανίχνευσης κοινοτήτων σε δίκτυο που θα χρησιμοποιήσουμε έτσι ώστε να δούμε πιο δουλεύει καλύτερα για το σκοπό μας, εξηγούμε πως ένα κείμενο μετατρέπεται σε διάνυσμα μέσω του μοντέλου BoW και πώς αυτό αξιοποιείται.

Έπειτα αναλύουμε το μοντέλο GoW και εξηγούμε μέσω ποιας διαδικασίας ένα κείμενο μπορεί να μετατραπεί και να γίνει αναπαράσταση αυτού σε γράφημα. Συνεχίζουμε με τα μέτρα αξιολόγησης των μεθόδων τα οποία χρησιμοποιούμε ώστε να αποφανθούμε για το ποια μέθοδος είναι πιο αποτελεσματική. Στη συνέχεια εφαρμόζουμε όλα αυτά τα μοντέλα και τα μέτρα σε πραγματικές συλλογές κειμένων ώστε να δούμε ποια μέθοδος έχει τα καλύτερα αποτελέσματα σε πραγματικά δεδομένα. Στο τέλος προσπαθούμε να κατασκευάσουμε τα αντίστοιχα μοντέλα για εικόνες, να δημιουργήσουμε γράφημα εικόνας με visual words και να αξιολογήσουμε μέσω πειραμάτων σε πραγματικά δεδομένα κατά πόσο αυτό το μοντέλο δουλεύει.

Τα μέτρα κεντρικότητας τα οποία μελετήσαμε παρουσιάζονται στο Κεφάλαιο 1. Πρώτον, η κεντρικότητα βαθμού (degree centrality) όπου είναι η κεντρικότητα που προκύπτει από το βαθμό του κάθε κόμβου, δηλαδή το πλήθος των συνδέσεων που έχει ένας κόμβος. Η κεντρικότητα βαθμού είναι ένας τοπολογικός δείκτης που αναδεικνύει το πόσο σημαντικός είναι ένας κόμβος σε ένα γράφημα. Το δεύτερο μέτρο που εξετάζουμε είναι η κεντρικότητα ενδιαμεσότητας (betweenness centrality) όπου βρίσκει τον πιο κεντρικό κόμβο με βάση τη θέση στο γράφημα, δηλαδή πόσο «ενδιάμεσος» είναι ο κόμβος, που εκτιμάται από το ποσοστό των μονοπατιών που διέρχονται από τον κόμβο αυτό. Ένα τρίτο μέτρο που μελετάμε είναι η κεντρικότητα εγγύτητας (closeness centrality) που υπολογίζεται ως το

μέσο μήκος των ελάχιστον μονοπατιών μεταξύ ενός συγκεκριμένου κόμβου και όλων των υπολοίπων μέσα στο δίκτυο έτσι ώστε να παίρνουν μεγαλύτερες τιμές οι πιο κεντρικές κορυφές με αποτέλεσμα να εκφράζει το πόσο κοντά είναι η συγκεκριμένη κορυφή στις υπόλοιπες. Τέταρτο μέτρο κεντρικότητας είναι η κεντρικότητα ιδιοδιανύσματος (eigenvector centrality), που είναι ένα μέτρο κεντρικότητας που συνυπολογίζει τόσο το πλήθος των γειτονικών κόμβων, όσο και τη σημαντικότητα του κάθε γείτονα. Κεντρικότητα PageRank είναι ένα μέτρο που εισήχθη στην βιβλιογραφία ως μέτρο που μετρά τη σημαντικότητα ιστοσελίδων στο διαδίκτυο και όπως και η κεντρικότητα ιδιοδιανύσματος βασίζεται στην σημαντικότητα όλων των γειτονικών κόμβων. Την ιδιότητα όμως να συνυπολογίζεται η σημαντικότητα των γειτονικών κόμβων την έχουν και οι κεντρικότητες mapping entropy (ME), mapping entropy betweenness (MEB) και mapping entropy closeness (MEC). Σε κάθε μια από τις προηγούμενες τρείς κεντρικότητες η σημαντικότητα των γειτόνων που λαμβάνεται υπόψιν είναι η κεντρικότητα βαθμού, η κεντρικότητα ενδιαμεσότητας και η κεντρικότητα εγγύτητας αντίστοιχα ώστε να πολλαπλασιάσει την αντίστοιχη κεντρικότητα (βαθμού, ενδιαμεσότητας, εγγύτητας) του κόμβου με ένα επιπλέον βάρος που είναι μια συνάρτηση εντροπίας. Επιπλέον, μελετούμε την κεντρικότητα coreness η οποία εξετάζει εάν ένας κόμβος ανήκει σε πλήρες υπογράφημα βαθμού k, ένα υπογράφημα δηλαδή του οποίου όλοι οι κόμβοι συνδέονται μεταξύ τους και έχουν βαθμό ακριβώς k. Η εκκεντρότητα (eccentricity) επίσης θεωρείται μέτρο κεντρικότητας διότι ποσοτικοποιεί πόσο μακριά είναι ένας κόμβος από τον κόμβο που απέχει περισσότερο από αυτόν μέσα στο γράφημα. Τέλος, εξετάζουμε τον συντελεστή συσταδοποίησης (clustering coefficient) ο οποίος εφαρμόζεται μεμονωμένα σε κάθε κόμβο του δικτύου και δείχνει την απόσταση που απέχουν οι γείτονες του κεντρικού κόμβου από τη δημιουργία κλίκας (πλήρες γράφημα).

Έπειτα περιγράφουμε και χρησιμοποιούμε τους αλγόριθμους ανίχνευσης κοινοτήτων Girvan-Newman, Fast greedy, Louvain, Infomap, Label propagation και walktrap σε γραφήματα. Στον αλγόριθμο τους οι Newman και Girvan προτείνουν πως σε ένα γράφο που περιέχει ομάδες κόμβων οι οποίες συνδέονται ασθενώς με λίγες ακμές, όλα τα ελάχιστα μονοπάτια μεταξύ αυτών των ομάδων πρέπει να περνάνε από μία από αυτές τις ακμές. Στον αλγόριθμο Label propagation προτείνονται τα εξής: Έστω ότι σε έναν γράφο

έχουμε έναν κόμβο ν με γείτονες τους κόμβους $\nu_1, \nu_2, \ldots, \nu_k$ όπου κάθε κόμβος έχει μια ιδιότητα η οποία χαρακτηρίζει την κοινότητα που ανήκει. Η κοινότητα του ν θα καθοριστεί από τις κοινότητες των γειτόνων του, καθώς σε κάθε επανάληψη του αλγορίθμου ο ν θα υιοθετεί την κοινότητα στην οποία ανήκουν οι περισσότεροι από τους γείτονές του. Καθώς οι ετικέτες διαμοιράζονται στον γράφο, σύντομα δημιουργούνται στενά συνδεδεμένες ομάδες κόμβων με κοινή ετικέτα. Ο αλγόριθμος Infomap είναι ένας από τους πιο αποτελεσματικούς αλγορίθμους ανίχνευσης κοινοτήτων. Συνδυάζει τεχνικές βασισμένες στην πληροφορία και τους τυχαίους περίπατους. Εξερευνά την τοπολογία του γράφου χρησιμοποιώντας έναν αριθμό από τυχαίους περίπατους συγκεκριμένου μήκους και μια δεδομένη πιθανότητα μεταβίβασης σε ένα τυχαίο κόμβο, ώστε να εντοπίσει το ελάχιστον μήκος κωδικοποίησης, διότι σε προηγούμενη εργασία τους οι Rosvall and Bergstrom έδειξαν ότι το πρόβλημα ανίχνευσης κοινοτήτων σε δίκτυο είναι ισοδύναμο με την ελαχιστοποίηση του μήκους κωδικοποίησης ενός τυχαίου περιπάτου σε αυτό. Ο αλγόριθμος Walktrap παράγει τυχαίους περιπάτους οι οποίοι είναι πιο πιθανό να εγκλωβιστούν μέσα σε μια κοινότητα σε σχέση με το να κάνουν μεταβάσεις από μια κοινότητα σε μια άλλη.

Στο Κεφάλαιο 2 παρουσιάζουμε το Bag of words (BoW) μοντέλο το οποίο βασίζεται στην αναπαράσταση κειμένου σε διάνυσμα χρησιμοποιώντας tf-idf scores τα οποία και αναλύουμε. Αναλύουμε ακόμη το πιο γενικά n-gram μοντέλα και παρουσιάζουμε μια λίστα λέξεων (stopwords) που αφαιρούνται κατά την επεξεργασία ενός κειμένου. Σε προβλήματα όπου απαιτείται η αναπαράσταση του κειμένου με το μοντέλο BoW κρατάμε μόνο τη ρίζα της κάθε λέξης. Ακολουθεί περιγραφει του μοντέλου Graph of words (GoW) όπου το κείμενο αναπαρίσταται σε μορφή γραφήματος με κόμβους τις λέξεις που συνδέονται μεταξύ τους αν ακολουθούνται σύμφωνα με το παράθυρο (πλήθος διαδοχικών λέξεων που συνδέονται μεταξύ τους) που έχουμε θέσει.

Στο μοντέλο Bag of words το κείμενο αναπαρίσταται σαν ένα σάκο που περιέχει όλες τις λέξεις του κειμένου ανεξαρτήτου γραμματικής. Το πλήθος των εμφανίσεων της λέξης στο κείμενο είναι γνωστό και σαν συχνότητα εμφάνισης (term frequency) της λέξης. Τα tf-scores είναι ένα στατιστικό νούμερο που μας δείχνει πόσο σημαντική είναι μια λέξη μέσα σε ένα κείμενο και με τη χρήση τους είναι δυνατόν να αναπαραστήσουμε το κείμενο με τη

μορφή διανύσματος και μπορούμε με αυτό τον τρόπο να συγκρίνουμε την ομοιότητα μεταξύ δύο κειμένων, χρησιμοποιώντας μια οποιαδήποτε απόσταση. Ακόμη μπορούμε να τα χρησιμοποιήσουμε στην ανάκτηση κειμένου από μια συλλογή κειμένων σύμφωνα με ένα ερώτημα που έχουμε θέσει. Εκτός από την ανάκτηση κειμένου το μοντέλο Bag of Words μπορεί να συμβάλει και στην κατηγοριοποίηση κειμένων, όπου η συχνότητα εμφάνισης της κάθε λέξης χρησιμοποιείται για την εκπαίδευση του μοντέλου μηχανικής μάθησης. Επιπλέον το μοντέλο Bag of words επιτρέπει την σύγκριση για ομοιότητα μεταξύ οποιονδήποτε δυο κειμένων σε προβλήματα συσταδοποίησης.

Εκτός όμως από το μοντέλο Bag of words, που χρησιμοποιεί μεμονωμένες τις λέξεις, το n-gram μοντέλο είναι μια επέκταση όπου χρησιμοποιεί ζευγάρια λέξεων (n=2), τριπλέτες λέξεων (n=3) και ούτω καθεξής. Χρησιμοποιώντας για παράδειγμα μόνο τη συχνότητα εμφάνισης λέξεων στο κείμενο το μοντέλο δε λαμβάνει υπόψιν το γεγονός ότι μετά από ένα όνομα ακολουθεί ρήμα στο κείμενο. Το n-gram μοντέλο μπορεί να φέρει αυτή την πληροφορία. Ένα πρόβλημα σε αυτή τη μέθοδο είναι ότι κάποιες λέξεις όπως τα άρθρα έχουν υψηλή συχνότητα εμφάνισης στο κείμενο χωρίς όμως να προσδίδουν κάποιο νόημα. Έτσι σε κάθε σύγκριση για ομοιότητα μεταξύ κειμένων αυτές οι λέξεις αφαιρούνται. Αυτές οι λέξεις λέγονται stopwords και για κάθε γλώσσα υπάρχει μία λίστα σύμφωνα με την οποία γίνεται η διαγραφή τους από τα κείμενα.

Το μοντέλο GoW είναι μια αναπαράσταση του κειμένου σε ένα γράφημα, όπου οι κόμβοι είναι οι λέξεις του κειμένου. Για ένα δοσμένο παράθυρο από N λέξεις, όλες οι λέξεις του παραθύρου συνδέονται και κάθε σύνδεσμος αναπαριστά τη συνεμφάνιση ενός ζεύγους λέξεων στο παράθυρο αυτό.

Στο Κεφάλαιο 3 εφαρμόζουμε το μοντέλο GoW στην εξόρυξη λέξεων-κλειδιά σε κείμενο και συγκρίνουμε με τα αποτελέσματα της εξόρυξης λέξεων κλειδιών από το μοντέλο BoW. Συγκρίνουμε την απόδοση του κάθε ενός απο τα δυο μοντέλα σε δυο δημόσια διαθέσιμες συλλογές και τις αξιολογούμε χρησιμοποιώντας μέτρα όπως ακρίβεια (precision@10), μέση ακρίβεια (mean Average Precision) και ο συντελεστής Jaccard (Jaccard coefficient).

Η σύγκριση έγινε χρησιμοποιώντας τα μέτρα κεντρικότητας και τις μεθόδους ανίχνευσης κοινοτήτων που παρουσιάζουμε, όταν αυτά εφαρμόζονται στο GoW για την εξόρυξη

λέξεων κλειδιών. Επιλέξαμε αυτές τις μεθόδους ως τις πιο γνωστές μεθόδους για την εύρεση κεντρικών κόμβων σε ένα γράφημα, ώστε να αναγνωρίσουμε το σύνολο λέξεων που περιγράφει πιο αντιπροσωπευτικά ένα αρχείο κειμένου. Τα μέτρα που επιλέξαμε για την αξιολόγηση τους έχουν χρησιμοποιηθεί στη βιβλιογραφία για την αξιολόγηση αποτελεσμάτων από προβλήματα ανάκτησης (retrieval problems) και είναι ικανά για τη σύγκριση ομοιότητας οποιονδήποτε δυο συλλογών λέξεων.

Οι μέθοδοι που επιλέξαμε για την σύγκριση ομαδοποιούνται σε δύο κατηγορίες. Αρχικά εφαρμόζονται τα μέτρα κεντρικότητας και τα πιο γενικά μέτρα που βασίζονται στις κεντρικότητες (transitivity, coreness, clustering coefficient) και στη συνέχεια κατατάσσουμε όλες τις λέξεις του κειμένου από την πιο κεντρική στη λιγότερο σύμφωνα με τη θέση της στην αναπαράσταση GoW. Έπειτα με τον αλγόριθμο ανίχνευσης κοινοτήτων στο GoW βρίσκουμε την μεγαλύτερη κοινότητα η οποία περιέχει τους κόμβους που είναι και οι λέξεις κλειδιά.

Επίσης εξετάζουμε και τρείς ακόμη κεντρικότητες στο πρόβλημα εξόρυξης λέξεων κλειδιά, α) mapping entropy, β) mapping entropy betweenness και γ) mapping entropy closeness το οποίο και εισάγουμε για πρώτη φορά στην εργασία αυτή.

Αντιθέτως με τη μέθοδο μεγιστοποίησης της modularity (Edge betweenness) που έχει χρησιμοποιηθεί για την εξαγωγή κοινοτήτων λέξεων κλειδιών, εξετάσαμε την εφαρμογή ανίχνευσης κοινοτήτων λέξεων κλειδιών στην αναπαράσταση GoW, οι οποίες εξάγονται με τις μεθόδους fast greedy, infomap, label propagation, Louvain και walktrap που αναλύουμε.

Για τα μέτρα αξιολόγησης ορίζουμε πρώτα ως $\mathcal{C}$ τη συλλογή των κειμένων και σημειώνουμε ως $\mathcal{R}$ το σύνολο των ανακτημένων αποτελεσμάτων που συμφωνούν με το ερώτημα q. Επίσης σημειώνουμε με $\mathcal{T}$ το σύνολο των σχετικών κειμένων, οπου οι φράσεις προέρχονται απο δοσμένο ground truth.

Ακρίβεια (Precision) ορίζεται ως συνάρτηση των ανακτημένων αποτελεσμάτων $\mathcal{R}$ απο τη συλλογή κειμένων $\mathcal{C}$ που είναι σχετικά με το ερώτημα q. Στην ακρίβεια χρησιμοποιούνται ολα τα ανακτημένα κείμενα, αλλά μπορεί ακόμη να υπολογιστεί με δοσμένο κατώφλι των

πρώτων n αποτελεσμάτων που επιστρέφονται απο το σύστημα και ονομάζεται ακρίβεια στα n, και συνήθως συμβολίζεται με *P@n*. Μέση ακρίβεια είναι ένα μέτρο που επηρεάζεται από τη σειρά των σχετικών εγγράφων. Δεν λαμβάνει υπόψη μόνο τον αριθμό των ανακτηθέντων εγγράφων που είναι συναφή, αλλά και τη θέση τους στην κατάταξη των αποτελεσμάτων που επιστράφηκαν. Δοσμένου πλήθους Q ερωτημάτων, ως mean Average precision (mAP) ορίζεται η μέση τιμή όλων των αποτελεσμάτων της μέσης τιμής για κάθε ερώτημα. Ανάκληση καλείται η συνάρτηση των σχετικών κειμένων που έχουν ανακτηθεί απο τη συλλογη $\mathcal{C}$ προς το συνολικό αριθμό των σχετικών κειμένων $\mathcal{T}$. F_1-*score* είναι το μέτρο που είναι ο αρμονικός μέσος της ακρίβειας και της ανάκλησης. Jaccard similarity χρησιμοποιείται στη στατιστική για τη σύγκριση ομοιότητας δυο δειγμάτων και ορίζεται ως το πλήθος της τομής τους ως προς το πλήθος της ένωσης τους. Στο πρόβλημα σύγκρισης δύο λιστών κειμένου, ο δείκτης ορίζεται ως το πλήθος των κοινών λέξεων ως προς το πλήθος των λέξεων που εμφανίζονται και στις δύο λίστες.

Σε δύο συλλογές κειμένων για τις οποίες γνωρίζουμε τις λέξεις κλειδιά που υπάρχουν σε κάθε κείμενο εφαρμόζουμε στατιστική αναπαράσταση του κειμένου με tf-idf scores και αναπαράσταση του κειμένου με γραφήματα χρησιμοποιώντας τα μοντέλα BoW και GoW αντίστοιχα. Στην περίπτωση της μιας συλλογής όπου ο λόγος είναι πιο δομημένος, η σειρά των λέξεων είναι σημαντική αφού οι συνδέσεις γίνονται μεταξύ λέξεων που βρίσκονται στο ίδιο παράθυρο. Άρα, το GoW υπερτερεί του BoW σε αυτή την περίπτωση.

Οι περιπτώσεις παραθύρων που παίρνουμε είναι δύο. Στην μια περίπτωση η κάθε λέξη συνδέεται με την επόμενη (N=2) και στην άλλη η κάθε λέξη συνδέεται με τις δυο επόμενες (N=3). Στη δεύτερη (N=3) τα αποτελέσματα είναι καλύτερα όμως αν συνεχίσει να αυξάνεται το μέγεθος του παραθύρου το γράφημα θα γίνεται πολύ πυκνό, οι κεντρικότητες δεν θα διαφέρουν και όλο το γράφημα θα έχει μια μόνο κοινότητα.

Ανάμεσα στα μέτρα κεντρικότητας και σε όλες τι μεθόδους που εξετάσαμε τα καλύτερα αποτελέσματα εμφανίζονται στην κεντρικότητα εγγύτητας και ακολουθούν οι κεντρικότητες MEB και MEC όταν η αξιολόγηση γίνεται με τον δείκτη Jaccard. Ανάμεσα στις μεθόδους ανίχνευσης κοινοτήτων η μέθοδος που ξεχωρίζει είναι η Infomap αλλά σε σχέση με τα υπόλοιπα μέτρα κεντρικότητας υποαποδίδει.

Στο Κεφάλαιο 4 παρουσιάζουμε τους πιο γνωστούς αλγόριθμους συσταδοποίησης οι οποίοι εφαρμόζονται και σε συσταδοποίηση κειμένου. Οι αλγόριθμοι αυτοί είναι οι DBSCAN, ιεραρχικής συσταδοποίησης (hierarchical clustering), k-means και Latent Dirichlet Allocation (LDA). Για την αξιολόγηση και τη σύγκριση των μεθόδων αυτών παρουσιάζουμε και εφαρμόζουμε τα μέτρα αξιολόγησης Normalized Mutual Information (NMI), Variation of Information (VI), δείκτης Rand και προσαρμοσμένος δείκτης Rand (Adjusted Rand).

Ο αλγόριθμος DBSCAN βασίζεται σε δυο παραμέτρους οι οποίοι ορίζονται από την πυκνότητα των συστάδων (ε) και από το ελάχιστο πλήθος των σημείων που μπορεί να περιέχει μια συστάδα ($MinPts$). Η μέθοδος δεν απαιτεί εκ των προτέρων την γνώση του πλήθους των συστάδων και είναι ικανή να εντοπίσει και να απομονώσει τον θόρυβο, δηλαδή σημεία τα οποία δεν ανήκουν σε καμία από τις συστάδες.

Η μέθοδος ιεραρχικής συσταδοποίησης (hierarchical clustering) έχει δυο προσεγγίσεις, μια διαχωριστική και μια συζευκτική. Στη διαχωριστική όλα τα σημεία ανήκουν σε μια συστάδα και με διαδοχικά βήματα γίνεται διαίρεση των ομάδων με κριτήριο την απόσταση όλων των σημείων μεταξύ τους, εως ότου κάθε σημείο να ανήκει στη δική του ξεχωριστή συστάδα. Στην αντίθετη περίπτωση ξεκινάμε με κάθε σημείο στη δική του συστάδα και με κριτήριο την απόσταση διαδοχικά δημιουργούνται ομάδες μέχρι όλα τα σημεία να ανήκουν σε μια μόνο συστάδα. Και στις δυο περιπτώσεις προκύπτει ένα δενδρόγραμμα το οποίο «κόβεται» σε κατάλληλο ύψος ώστε να προκύψουν οι αντίστοιχες συστάδες.

Στη μέθοδο k-means clustering απαιτείται η εισαγωγή του πλήθους των συστάδων και ο τυχαίος ορισμός των αρχικών συστάδων. Με μια επαναληπτική διαδικασία κάθε σημείο αντιστοιχίζεται σε μια συστάδα, το κέντρο της οποίας επαναπροσδιορίζεται με στόχο την ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος εντός των συστάδων για όλες τις συστάδες.

Το στατιστικό μοντέλο LDA συσταδοποιεί μια συλλογή από αρχεία κειμένου σε θέματα. Κάθε συστάδα είναι μια συλλογή από αρχεία κειμένου που περιγράφουν ένα θέμα. Κάθε θέμα έχει μια αναπαράσταση ως πολυωνυμική κατανομή πιθανότητας για την εκτίμηση των παραμέτρων της οποίας αξιοποιείται ότι η δεσμευμένη εκ των προτέρων κατανομή της πολυωνυμικής είναι η κατανομή Dirichlet.

Τα μέτρα αξιολόγησης του αποτελέσματος της συσταδοποίησης χωρίζονται α) σε μέτρα που ελέγχουν ζευγάρια σημείων (Rand, Adjusted Rand) και β) σε αυτά που βασίζονται στην πληροφορία των διαμερίσεων (NMI, VI). Στους δείκτες Rand και Adjusted Rand εξετάζεται αν τα δυο μέλη ενός ζευγαριού σημείων ανήκουν στην ίδια ή σε διαφορετικές συστάδες εξετάζοντας τις δυο διαφορετικές διαμερίσεις. Ο δείκτης NMI είναι η κανονικοποιημένη αμοιβαία πληροφορία των δυο διαμερίσεων (η διαμέριση που επισημαίνεται από το σύνολο δεδομένων και η διαμέριση που προκύπτει από τον αλγόριθμο) και VI είναι η πληροφοριακή απόσταση των δυο διαμερίσεων. Όσο μικρότερος είναι ο δείκτης VI και όσο μεγαλύτεροι οι υπόλοιποι δείκτες τόσο τα αποτελέσματα γίνονται περισσότερο αξιόλογα.

Εφαρμόζουμε τις τεχνικές συσταδοποίησης κειμένου σε δημόσια διαθέσιμες συλλογές κειμένων από ειδήσεις αλλά και αρχείων κειμένων χρηστών από το Twitter, όπου είναι γνωστό εκ των προτέρων σε ποιο θέμα ανήκει το κάθε κείμενο και ποιο είναι το πλήθος των θεμάτων.

Συμπεραίνουμε ότι η μέθοδος LDA δίνει καλύτερα αποτελέσματα σε δυο από τις τρείς συλλογές κειμένων ως προς τους δείκτες NMI, VI και Adjusted Rand. Στην άλλη συλλογή καλύτερα αποτελέσματα έχει η εφαρμογή του hierarchical clustering όπου χρησιμοποιήθηκε το βέλτιστο ύψος του δενδρογράμματος σύμφωνα με το δείκτη NMI, εξετάζοντας όλες τις δυνατές τιμές του.

Ο δείκτης Rand μεγιστοποιείται σε κάθε συλλογή στην περίπτωση του hierarchical clustering, γεγονός που προκύπτει λαμβάνοντας υπόψιν ότι ο δείκτης Rand προκύπτει από ζευγάρια αντικειμένων μέσα από μια συλλογή και ταυτόχρονα ότι η μέθοδος hierarchical clustering ενώνει κοντινά αντικείμενα σε ζευγάρια, χωρίς αυτό να είναι ένας γενικός κανόνας.

Στην περίπτωση των αναρτήσεων στο Twitter, η συσταδοποίηση τους είναι πιο αποτελεσματική στην περίπτωση, όπου αξιοποιείται η θεωρία γραφημάτων και, επιπλέον, είναι δυνατή η οπτικοποίηση της συλλογής και των αποτελεσμάτων, με στόχο την ποιοτική αξιολόγηση.

Στο Κεφάλαιο 5 παρουσιάζουμε το μοντέλο Bag of Visual Words (BoVW) που αποτελεί τη γενίκευση του μοντέλου BoW στην αναπαράσταση μιας εικόνας σε διάνυσμα, όπως γίνεται η αναπαράσταση ενός κειμένου σε διάνυσμα χρησιμοποιώντας ένα κοινό λεξικό από τις λέξεις που εμφανίζονται στη συλλογή και προτείνουμε το μοντέλο αναπαράστασης εικόνας με γράφημα Graph of Visual Words (GoVW).

Στο μοντέλο Bag of Visual Words (BoVW) το λεξικό από «οπτικές λέξεις» (visual words) δημιουργείται συσταδοποιώντας με τη μέθοδο k-means τα διανύσματα SIFT, τα οποία προκύπτουν, μέσω εργαλείων, από τα κρίσιμα σημεία της κάθε εικόνας εκεί δηλαδή όπου εμφανίζονται απότομες αλλαγές στο σχήμα (γωνίες, ακμές κλπ). Κάθε εικόνα έχει ένα διαφορετικό αριθμό από κρίσιμα σημεία τα οποία οδηγούν σε διανύσματα 128 διαστάσεων.

Επιπλέον, προτείνουμε το μοντέλο αναπαράστασης εικόνας με γράφημα Graph of Visual Words (GoVW) όπου αντικαθιστούμε την συσταδοποίηση k-means στο μοντέλο BoVW με τη δημιουργία ενός γραφήματος, όπου δυο διανύσματα SIFT συνδέονται αν η μεταξύ τους απόσταση είναι μικρότερη από το πρώτο τεταρτημόριο των αποστάσεων τους:

$$l(s_\kappa, s_\lambda) = \begin{cases} 1 & \alpha\nu\ d(s_\kappa, s_\lambda) < \varepsilon \\ 0 & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$$

όπου (s_κ, s_λ) ένα τυχαίο ζεύγος από διανύσματα SIFT, ε το πρώτο τεταρτημόριο των αποστάσεων τους και $d(s_\kappa, s_\lambda)$ η μεταξύ τους απόσταση. Στο γράφημα που προκύπτει εφαρμόζουμε αλγόριθμο ανίχνευσης κοινοτήτων με στόχο τη δημιουργία ομάδων (συστάδων) SIFT διανυσμάτων, δηλαδή οπτικές λέξεις. Ένα από τα πλεονεκτήματα της μεθόδου αυτής είναι ότι δεν απαιτεί την εκ των προτέρων γνώση του πλήθος των οπτικών λέξεων (συστάδων).

Εφαρμόζουμε τα μοντέλα BoVW και GoVW σε δυο δημόσια διαθέσιμες συλλογές εικόνων και παρατηρούμε ότι ο δείκτης NMI αυξάνεται κατά 6.68% και 16.11% στις δυο συλλογές αντίστοιχα χρησιμοποιώντας τη μέθοδο GoVW έναντι της BoVW. Όσον αφορά το δείκτη Adjusted Rand τη μια φορά αποδίδει στο ένα μοντέλο και την επόμενη στο άλλο.

Εκτός από την ποσοτική αξιολόγηση των δυο μεθόδων, παρουσιάζουμε και την ποιοτική, όπου δείχνουμε το αποτέλεσμα της ομαδοποίησης των εικόνων μια φορά όπως προκύπτει από το μοντέλο BoVW και μια από το μοντέλο GoVW. Παρατηρούμε ότι οι εικόνες της συλλογής που δείχνουν το ίδιο ή συναφές αντικείμενο και τείνουν να ανήκουν στην ίδια ομάδα (συστάδα) ομαδοποιήθηκαν με την μέθοδο GoVW, ενώ το αποτέλεσμα της μεθόδου BoVW τείνει να είναι περισσότερο ανομοιογενές.

ACKNOWLEDGEMENTS

I would like to thank the lecturers of the inter-faculty master program in Complex Systems and Networks of the Aristotle University of Thessaloniki for providing me solid background in network theory and applications, during the postgraduate courses.

This thesis would not have been completed without the support of the supervisory committee constituted from the professor Ioannis Antoniou, the senior researcher Dr. Stefanos Vrochidis and the postdoctoral researcher Dr. Ilias Gialampoukidis.

Last but not least, I would like to thank my parents for truly supporting me during my undergraduate and postgraduate studies.

PROLOGUE

Nowadays, multimedia are all around us (smartphones, WWW, social media, etc.), involving large streams of information that we need to handle efficiently and quickly. Multimedia usually appear in image representations, associated with text descriptions and/or textual tags or concepts. In order to handle big collections of text documents or image collections, it is necessary to first cluster them into groups of similar objects. At the level of a single document, reading lengthy text documents is a time consuming process that needs to be assisted by a keyword extraction mechanism, in order to provide the reader a quick overview of the main topics of the text document.

In chapter 1, we present the necessary background in graph centralities and community detection approaches, which will be used in text representation and in keyword extraction. The most prominent centrality measures and other centrality measures, such as coreness, eccentricity and clustering coefficient are defined and community detection approaches are reported, aiming to find the most central community in the graph of words of a document.

In chapter 2, the Bag-of-Words (BoW) and the Graph-of-Words (GoW) models are presented. The BoW model is based on a vector representation of a text document using tf-idf scores. Additionally, the n-gram model representation is also reported, as well as the list of stopwords that are removed when text documents are processed, before stemming. The Graph-of-Words model follows, in which each text document is represented as a graph, where nodes are words and links are added according to the co-occurrence of two words in a window of N successive words.

In chapter 3, we apply the BoW and the GoW models in the keyword extraction problem. We compare their performance in two publicly available datasets using the evaluation measures Precision@10, mean Average Precision and Jaccard coefficient. The comparison is done using the centrality measures and communities, presented in Chapter 1. We selected these methods as the most prominent methods to identify central nodes in a graph.

In chapter 4, we present popular clustering approaches that have also been used in the context of text clustering. Firstly, a density-based algorithm called DBSCAN is reported and secondly, k-means clustering is presented. Hierarchical clustering is thirdly presented and finally we refer to Latent Dirichlet Allocation as a well performing method in topic modeling. Evaluation measures in clustering are also described, such as Normalized Mutual Information, Variation of Information, Rand and adjusted Rand indices. Finally, we examine which of the clustering approaches perform better in three public datasets of news articles and in Twitter posts, which were collected according to five popular hashtags, in order to also examine short text clustering, with respect to the considered evaluation measures. Qualitative evaluation also appears in Section 4.4, where a graph of Twitter posts is created, using the Jaccard similarity, and the result of the community detection algorithms provide the final clustering.

In chapter 5, we present the Bag of Visual Words (BoVW) model that has been used to represent an image using a statistical approach, similar to the BoW representation in text modelling. We also propose an alternative approach for the creation of visual words, but using a graph model, combined with a community detection approach on the formulated graph of SIFT descriptors, namely the Graph of Visual Words (GoVW). We evaluate our proposed model in two public datasets of image collections, which provide ground truth annotation for clustering purposes, and we find evidence that our method is more efficient than the BoVW model in the image clustering task.

Chapter 1. Graph centralities and community detection

In this chapter we present the necessary background in graph centralities and community detection approaches, which will be used in text representation and in knowledge extraction from text, such as the keyword extraction problem. In Section 1.1, the most prominent centrality measures are defined and other centrality measures, such as coreness, eccentricity and clustering coefficient, follows in Section 1.2. Moreover, community detection approaches are reported in Section 1.3, in order to extract groups of nodes that are densely connected, aiming to find the most central community in the graph of words of a text document.

1.1 Centrality measures

In this section we present well-known centrality measures of a graph, such as degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, page rank

centrality, mapping entropy centrality and its extensions, namely mapping entropy betweenness and mapping entropy closeness.

Given an undirected network $G(N, L)$ with N nodes and L links, the adjacency matrix $\boldsymbol{A}$ of a network $G(N, L)$ is a square matrix which is defined as follows:

$$A(n_i, n_j) = A_{ij} = \begin{cases} 1, if\ n_i, n_j\ are\ connected \\ 0, otherwise \end{cases} \quad (1.1)$$

In general, we denote by M_{ij} the (i, j) element of a matrix M.

Degree centrality

Degree of a node n_k, $\deg(n_k)$ is the number of edges connected to it. The maximum number of nodes that node n_k can be connected is $N - 1$ and the degree centrality (DC) of node n_k is defined as (Freeman, 1979):

$$DC_k = \frac{\deg(n_k)}{N - 1} \quad (1.2)$$

Betweenness centrality

Let n_i, n_j be two nodes and g_{ij} the number of geodesics (the shortest path between two nodes) linking n_i with n_j, then the probability of using one of them is $\frac{1}{g_{ij}}$. Let also $g_{ij}(n_k)$ the number of geodesics linking n_i and n_j that contain n_k then $\frac{g_{ij}(n_k)}{g_{ij}}$ is the probability to have a geodesic from n_i to n_j that passes through n_k. For all unordered pairs of nodes where $i \neq j \neq k$, the betweenness centrality (BC) of node n_k is defined as follows:

$$BC_k = \sum_{i<j}^{N} \frac{g_{ij}(n_k)}{g_{ij}} \quad (1.3)$$

where N is the number of nodes in the graph. But like degree centrality we need a measure from which the effect of network size has been removed. The maximum value for betweenness centrality in n_k is achieved only by the central node in a star graph (Freeman,1977), and it is:

$$1 + 2 + \cdots + N = \frac{N(N+1)}{2} \tag{1.4}$$

so all paths for all (i, j) pairs are:

$$1 + 2 + \cdots + N - 1 = \frac{N(N-1)}{2} \tag{1.5}$$

The central node is not contained in these paths, so we extract $N - 1$:

$$\frac{N(N-1)}{2} - (N-1) = \frac{N^2 - N}{2} - \frac{2(N-1)}{2} = \frac{N^2 - N - 2N + 2}{2} = \frac{N^2 - 3N + 2}{2} \tag{1.6}$$

Hence, the maximum of the betweenness centrality of node n_k is used to normalize the betweenness centrality of a node in the interval $[0,1]$, as referred in (Freeman,1977):

$$BC_k = \frac{2\sum_{i<j}^{N} \frac{g_{ij}(n_k)}{g_{ij}}}{N^2 - 3N + 2} \tag{1.7}$$

The fact that the normalized BC is in $[0,1]$, inherits the properties of a probability measure. Another centrality that is based on the geodesics of the graph is the closeness centrality.

Closeness centrality

Let $d(n_i, n_k)$ be the number of edges in the geodesic linking n_i and n_j. The sum of all distances from node n_k to all other nodes shows how far is the node n_k from all other nodes $n_i, i = 1,2, \dots, N$, on average:

$$Decentrality(farness) = \sum_{i=1}^{N} d(n_i, n_k) \tag{1.8}$$

Then the inverse of the decentrality of a node n_k (Sabidussi, 1966) shows how close is a node n_k from all other nodes:

$$closeness = \frac{1}{\text{decentrality}} = \frac{1}{\sum_{i=1}^{N} d(n_i, n_k)} \tag{1.9}$$

The corresponding measure from which the effect of network size has been removed is:

$$Decentrality(farness) = \frac{\sum_{i=1}^{N} d(n_i, n_k)}{N-1} \tag{1.10}$$

and the closeness centrality CC of the node n_k is defined as:

$$CC_k = \frac{N-1}{\sum_{i=1}^{N} d(n_i, n_k)} \tag{1.11}$$

The centrality of all neighbors puts weights on the centrality of a node, defining the eigenvector centrality below.

Eigenvector centrality

Let $x = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$ be a vector where x_k the centrality of node n_k. If we want count the centrality x_k of node n_k which is dependent of n_k's network neighbours centrality (but not of their number) then

$$x_k = \frac{1}{\lambda} \sum_{j=1}^{N} A_{kj}\, x_j \tag{1.12}$$

where λ is a constant. Therefore, the vector x is:

$$x = \frac{1}{\lambda} Ax \Leftrightarrow \lambda x = A \tag{1.13}$$

and hence we can solve the eigenvector problem of adjacency matrix A. For

$$x_k \geq 0, \forall\, k = 1,2, \dots N \tag{1.14}$$

It can be shown (using Perron-Frobenius theorem) (Gantmacher, 1998) that λ must be the largest eigenvalue of the adjacency matrix.

Similarly to the eigenvector centrality, the PageRank is based on the centrality of the node's neighborhood.

Page Rank

PageRank (PR) is a centrality measure, originally known as Google's algorithm (Brin and Page, 2012) which has been introduced to count how important is a Web page, and it is defined for node n_k as:

$$PR_k = \frac{1-d}{N} + d \sum_{n_i \in \mathcal{N}(n_k)} \frac{PR_i}{L(n_i)} \tag{1.15}$$

where d is the damping factor, typically set to 0.085, $L(n_i)$ is the number of links to node n_i and $\mathcal{N}(n_k)$ is the neighborhood of n_k, i.e. the set of nodes connected to node n_k.

In addition to the eigenvector centrality and the PageRank, a centrality measure has been introduced, also based on the centrality scores of all node's neighbors.

Mapping entropy

The set of nodes connected to node n_k, $\mathcal{N}(n_k)$ has been used to define the mapping entropy (ME) centrality which has recently been proposed in (Nie et al., 2016) as a function of the degree centrality:

$$ME_k = -DC_k \sum_{n_i \in \mathcal{N}(n_k)} \log DC_i \tag{1.16}$$

Mapping entropy is in fact the degree centrality DC_k weighted by the average *Shannon information* in the neighborhood of node n_k, where *Shannon information* is defined (Cover and Thomas, 2012) as:

$$\mathcal{I} = -\sum_k p_k \log p_k \tag{1.17}$$

for any probability distribution: $0 \leq p_k \leq 1$ and

$$\sum_k p_k = 1 \tag{1.18}$$

The normalization of centrality measures in [0,1] allows for considering the values as probabilities and, therefore, the Mapping Entropy has been generalized to the so called Mapping Entropy Betweenness, which is bases on the BC of all neighbors of a node.

Mapping entropy Betweenness

Mapping Entropy has been extended (Gialampoukidis et al., 2016a) by replacing the degree centrality with the Betweenness centrality, as follows:

$$MEB_k = -BC_k \sum_{n_i \in \mathcal{N}(n_k)} \log BC_i \tag{1.19}$$

Mapping Entropy Betweenness has been used in the context of key player identification in terrorism-related social media networks, constructed by the Twitter mentions from one user to another.

Mapping entropy extensions

In this thesis, we also examine whether other extensions of Mapping Entropy centrality are effective or not, in the context of unsupervised graph-based keyword extraction, such as the Mapping Entropy Closeness (MEC) centrality of node n_k :

$$MEC_k = -CC_k \sum_{n_i \in \mathcal{N}(n_k)} \log CC_i \tag{1.20}$$

Some other combinations of Mapping Entropy centrality which we construct in order to examine them in the keyword extraction problem from text documents, are:

CC which is weighted by the entropy of the nodes neighborhood based on the DC:

$$MECD_k = -CC_k \sum_{n_i \in \mathcal{N}(n_k)} \log DC_i \tag{1.21}$$

On the other hand, DC which is weighted by the entropy of the nodes neighborhood based on the CC:

$$MEDC_k = -DC_k \sum_{n_i \in \mathcal{N}(n_k)} \log CC_i \tag{1.22}$$

1.2 Other graph-based measures

Other measures that we use to identify the most central nodes of a graph are the coreness, the eccentricity and the local transitivity, known also as the (local) clustering coefficient.

Coreness

The k-core of a graph G is defined as the maximum subgraph of G in which all nodes have at least degree k. The coreness of a node of the graph G is k if it belongs to the k-core but not to the $(k+1)$-core.

Coreness has been used in the keyword extraction problem in order to find a group of words that is the most representative in a text document (Rousseau and Vazirgiannis, 2013).

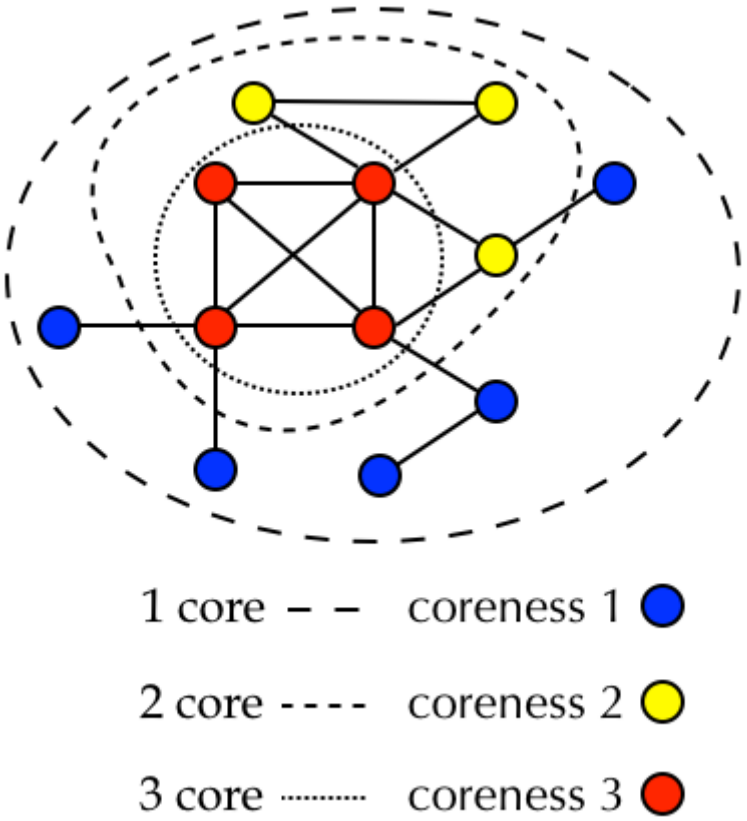


Figure 1.1[1]: The coreness of each node of an illustrative graph

Eccentricity

The eccentricity of a node k in a graph G is the greatest geodesic distance between the node k and any other node. It can be thought of as how far is a node from the node most distant from it in the graph. The geodesic between two nodes n_i, n_j is the shortest path linking n_i and n_j, as already presented in the definition of betweenness centrality in Section 1.1.

[1] https://chaoslikehome.wordpress.com/tag/k-shell/

The eccentricity can be considered as a centrality measure because the most central node of a graph has the minimum eccentricity. The minimum and the maximum eccentricity of a graph is called the radius and the diameter of the graph respectively.

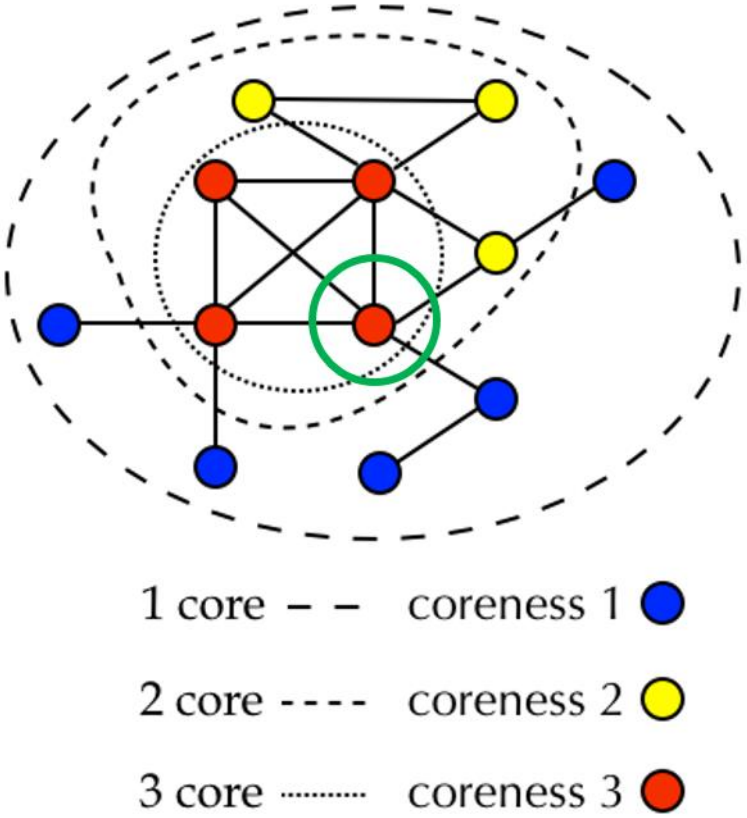


Figure 1.2: The selected node (in the green circle) is the node with the minimum eccentricity in the graph of Figure 1.1

Clustering coefficient (local transitivity)

The local clustering coefficient of a node n_i in a graph G quantifies how close the neighbors of n_i are to being a clique (complete graph). The local clustering coefficient of a node n_i in a directed graph is defined as:

$$C_i = \frac{|\{e_{jk}: n_j, n_k \in N_{i,} e_{jk} \in E\}|}{k_i(k_i - 1)} \quad (1.23)$$

where e_{jk} is the link from node n_j to n_k, N_i is the set of neighbours of n_i and k_i is the degree of node n_i.

In an undirected graph the formula is modified as follows:

$$C_i = \frac{2|\{e_{jk}: u_j, u_k \in N_{i,} e_{jk} \in E\}|}{k_i(k_i - 1)} \tag{1.24}$$

1.3 Community detection

In this section we present selected graph clustering approaches into communities of nodes, which can be used for large networks, in contradiction to the Girvan-Newman maximization of modularity, presented below. We describe the Louvain method, the Infomap minimization of codelength, the Walktrap, Fast greedy and Label propagation methods.

Girvan–Newman algorithm

The Girvan–Newman community detection algorithm (Girvan and Newman, 2002; Newman and Girvan, 2004) is a divisive hierarchical process, based on the edge betweenness centrality measure (Freeman, 1977), which is calculated following Brandes (Brandes, 2001) for a faster implementation. The edge betweenness determines the edges which are more possible to link different communities. The edge with the highest edge betweenness is removed and the other edges are re-assigned new edge betweenness scores. The process generates a dendrogram with root node the whole graph and leaves the graph vertices. In order to extract the detected communities, the modularity score is computed at each dendrogram cut, so as to be maximized. The modularity has been defined as follows (Newman and Girvan, 2004):

$$Q = \sum_i (e_{ii} - a_i^2), \qquad a_i = \sum_j e_{ij} \tag{1.25}$$

where e_{ij} are the elements of a $k \times k$ summetric matrix and k is the number of communities at which the graph is partitioned. The elements e_{ij} are defined as the fraction of all edges in the network that link vertices in community i to vertices in community j.

Modularity Maximization – Fast Greedy algorithm

The Girvan–Newman algorithm requires the maximization of a modularity function, as a stopping criterion, for the optimal extraction of communities. However, in (Clauset et al., 2004) an alternative hierarchical approach for community detection has been presented, using the modularity function as an objective function to optimize. Initially, all nodes are separate communities and any two communities are merged if the modularity increases. The algorithm stops when the modularity is not increasing anymore. The modularity function is defined as (Clauset et al., 2004):

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(i,j) \tag{1.26}$$

where A_{ij} is the (i,j) element of the adjacency matrix, m is the number of edges in the graph, k_i is the degree of node i and $\delta(i,j)$ is 1 if $i = j$ and 0 otherwise.

The modularity Q may also be generalized to any null model with expected number of edges between vertices i and j (Fortunato, 2010):

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - P_{ij}\right]\delta(i,j) \tag{1.27}$$

The modularity maximization algorithm of (Clauset et al., 2004) is a faster method to detect communities based on the modularity maximization, as proposed in the Girvan–Newman community detection algorithm.

Louvain method

The Louvain method (Blondel et al., 2008) is based on the maximization of the modularity Q and involves two phases that are repeated iteratively. In the first phase, each node forms a community and for each node i the gain of modularity is calculated for removing vertex i from its own community and placing it into the community of each neighbor j of i. The vertex i is moved to the community for which the gain in modularity becomes maximal. In case the modularity decreases or remains the same, vertex i does not change community. The first phase is completed when the modularity cannot be further increased. In the second phase, the detected communities formulate a new network with weights of the links between the new nodes being the sum of weights of the links between nodes in the corresponding two communities. In this new network, self-loops are allowed, representing links between vertices of the same community. At the end of the second phase, the first phase is re-applied to the new network, until no more communities are merged and the modularity attains its maximum.

Infomap method

The Infomap method (Rosvall and Bergstrom, 2008; Rosvall et al., 2010) is an information-theoretic approach for community detection. The inventors of the Infomap method are based on their previous work (Rosvall and Bergstrom, 2007), in which they showed that the problem of finding a community structure in networks is equivalent to solving a coding problem. In general, the goal of a coding problem is to minimize the information required for the transmission of a message. Initially, Infomap employs the Huffman code (Huffman, 1952) in order to give a unique name (codeword) in every node in the network.

In contrast to the Louvain method, Infomap minimizes the Shannon information (Cover and Thomas, 2012) required to describe the trajectory of a random walk on the network. A global information minimum (in bits) for the description of the trajectory of a random walk X on the network, with n states and corresponding probabilities p_i, is given by Shannon's source coding theorem (Cover and Thomas, 2012):

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{1.28}$$

which is the Shannon information of the random walk X.

The objective function, which minimizes the description length of a random walk on the network (described by the corresponding sequence of codewords on each visited node), is called the "map equation" (Rosvall and Bergstrom, 2008; Rosvall et al., 2010), and is minimized over all possible network partitions $\mathbf{M}$:

$$L(\mathbf{M}) = q_{\curvearrowright} H(\mathcal{L}) + \sum_{i=1}^{m} p_{\circlearrowright}^{i} H(\mathcal{P}^i) \tag{1.29}$$

The first term of Equation describes the entropy of the random walk movements between communities and the second part is the entropy of movements within communities (exiting the community i is considered a movement of the second term). The fraction of transitions within the i-th community is denoted by $p_{\circlearrowright}^{i}$ and $H(\mathcal{P}^i)$ is the entropy of community $\mathcal{P}^i$. The probability $q_{\curvearrowright}$ that the random walk switches communities on any given step is:

$$q_{\curvearrowright} = 1 - \sum_{i=1}^{m} p_{\circlearrowright}^{i} \tag{1.30}$$

The computational procedure followed for the minimization of $L(\mathbf{M})$ is presented in the supporting Appendix of (Rosvall and Bergstrom, 2008).

Other Methods

The **Label Propagation** method (Raghavan et al., 2007) initializes every node with a unique label and at each step every node adopts the label that most of its neighbors currently have. Hence, an iterative process is defined, in which densely connected groups of nodes form a consensus on a label and communities are extracted. The **Walktrap** method (Pons and Latapy, 2005) generates random short walks on the graph by simulating transitions between

nodes. Since short random walks tend to stay within the same community, it is possible to detect communities using such random walks.

Other methods for community detection involve density-based approaches, which are able to identify noise, i.e. nodes that do not belong to any of the communities. DBSCAN* (Campello et al., 2013), the graph analogue of the well-established DBSCAN algorithm (Ester et al., 1996), is such a density-based approach that could be applied to community detection. Similarly to DBSCAN, it relies on two parameters: the density level ε and a lower bound $MinPts$ for the number of nodes that may form a community. However, the estimation of these parameters is not trivial, so in order to address this issue, and in particular the estimation of $MinPts$, a **DBSCAN*-Martingale** approach has been proposed (Gialampoukidis et al., 2016b), which involves the construction of a Martingale process.

Chapter 2. Bag of Words and Graph of Words

The Bag-of-Words model is presented in this chapter. It is based on a vector representation of a text document using tf-idf scores, which are presented in Section 2.1. In this section, the more general n-gram model representation is also reported, as well as the list of stopwords that are removed when text documents are processed, before stemming. The Graph-of-Words model follows, in which text is represented as a graph, where nodes are words and links are added according to the co-occurrence of two words in a window of N successive words.

2.1 Bag of Words

The Bag-of-words (BoW) model is a representation which used in Natural Language Processing (NLP) and in Information Retrieval (IR). In this model, text is represented as a bag which contains all text's words free from grammar and word order. Word's multiplicity is the number of occurrences of a word in a document, known also as *term frequency* (tf).

Tf-idf scores and text retrieval

Term frequency (tf) scores are weighted by the inverse document frequency, to put less weight in words that appear in many documents. The tfidf scores are defined as:

$$tfidf_{ij} = \frac{n_{id}}{n_d} \log \frac{N}{n_i} \tag{2.1}$$

where $\frac{n_{id}}{n_d}$ is the term frequency and:

n_{id} = the number of occurrences of word i in document d

n_d = the number of words in document d

n_i = the number of occurrences of word i in the whole database

N = the total number of documents in the database

We note that the inverse document frequency of a word is given by the Shannon Information (Cover and Thomas, 2012):

$$\log \frac{N}{n_i} = i(word_i) = -\log \frac{|\{d \in D: word_i \in d\}|}{N} \tag{2.2}$$

Using tf-idf scores it is possible to represent text documents as vectors and to compare the similarity between any pair of documents. Any function of tf and idf scores can be used to retrieve text documents, in response to a query q. For example, the Okapi BM25 model (Robertson and Zaragoza, 2009) is a BoW retrieval function, which have been used by search engines to rank matching documents according to their relevance to a given query. Given a query q that contains the keywords $q_1, q_2, \dots, q_n$, the BM25 score of a document D is:

$$score(D,q) = \sum_{i=1}^{n} idf(q_i) \frac{f(q_i,D)(k_1+1)}{f(q_i,D) + k_1(1 - b + b\frac{|D|}{avgdl}} \tag{2.3}$$

where, $f(q_i, D)$ is the term's frequency defined as the number of times the query term q_i appears in the document D and $|D|$ is the length of the document D in words (terms). Moreover, $avgdl$ is the average document length over all the documents of the collection, k_1 and b are free parameters, usually chosen as $k_1 = 2$ and $b = 0.75$, and finally, $idf(q_i)$ is the inverse document frequency weight of the query term q_i, computed as:

$$idf(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \quad (2.4)$$

where, N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing q_i.

Apart from text retrieval, the BoW model is also involved in document classification methods, where the frequency of word appearance is used to train a classifier. Moreover, BoW allows for comparing the similarity between any two text documents for clustering problems. In the following, we present an example of a BoW representation of the collection with the two documents:

(A) John likes to play football. George likes football too.

(B) John also likes to read books.

From the documents (A) and (B), the set of all unique words is extracted, so as to create a vocabulary over the whole collection:

"John" "likes" "to" "play" "football" "George" "too" "also" "read" "books"

The term-frequency vectors of sentences (A) and (B) are:

(A) [1,2,1,1,2,1,1,0,0,0]
(B) [1,1,1,0,0,0,0,1,1,1]

and the corresponding tf-idf scores are:

(A) [0,-0.035,0,0.030,0,0.030,0.030,0,0,0]
(B) [0,-0.017,0.030,0,0,0,0,0.030,0.030,0.030]

Each number of the term frequency vectors is the multiplicity of the corresponding word. For example in vector (A) the first number is "1" which means that in the first sentence the word "John" appears one time. The second number is "2" which means that the second word "likes" appears twice in the first sentence. This vector is independent of the order of the words in the original document.

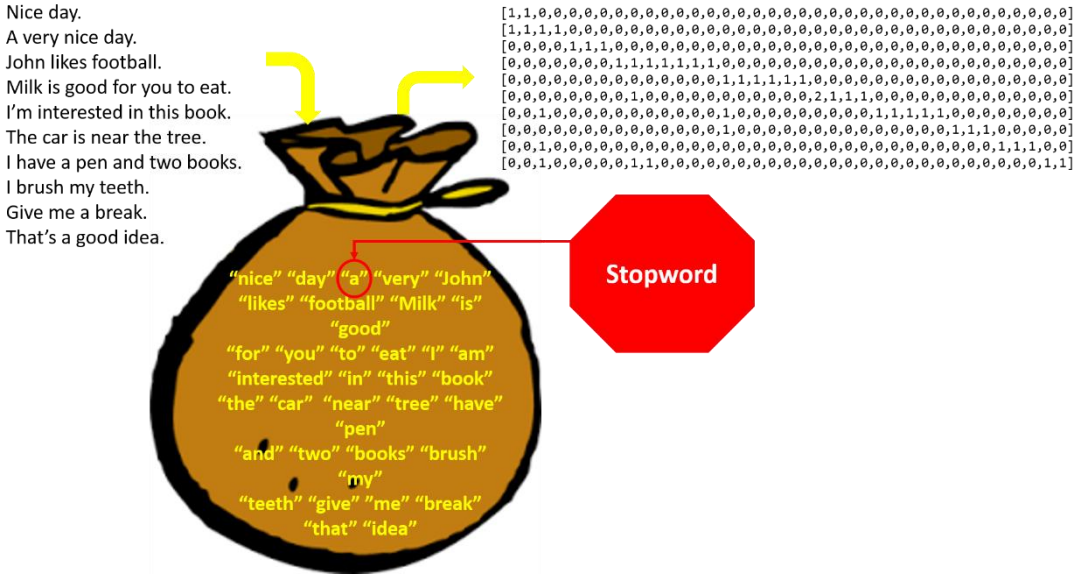


Figure 2.1: The BoW model for a list of sentences and the corresponding term frequency (tf) vectors. The tf-idf scores are presented in Appendix A.

However, it is possible to consider pairs, triplets or n-tuples of words as "terms", known as word n-grams.

N-gram model

Apart from the BoW formulation using single words, the n-gram model extends the use of single terms to pairs of words, triplets of words, etc. The n-gram model has as special case the BoW formulation using single words for $n = 1$.

Using for example only the word frequency in a text (unigrams), the model will not reveal the fact that after a name follows a verb in the text. The n-gram model can be used to return this information, as shown in the following bigram example:

"John likes" "likes to" "to play" "play football" "football George" "George likes" "likes football" "football too"

Stopwords

The problem of this method is that some words like articles "a", "the" etc. have the highest term frequency in the text, while they do not provide content for the document. In any comparison of the similarities among text documents, these words are removed so as to assist any clustering, classification or retrieval problem. These words are called "stopwords" and there is no universal list of stopwords per language. However, some lists have been created and are extensively used, such as the "SMART[2]" stopwords list and the list of English stopwords in the "tm[3]" package in R[4]. The SMART stopword list is presented in Table 2.1.

A	came	from	keep	one	sometime	unto
a's	can	further	keeps	ones	sometimes	up
able	can't	furthermore	kept	only	somewhat	upon
about	cannot	g	know	onto	somewhere	us
above	cant	get	knows	or	soon	use
according	cause	gets	known	other	sorry	used
accordingly	causes	getting	l	others	specified	useful
across	certain	given	last	otherwise	specify	uses
actually	certainly	gives	lately	ought	specifying	using
after	changes	go	later	our	still	usually
afterwards	clearly	goes	latter	ours	sub	uucp
again	Co	going	latterly	ourselves	such	v
against	com	gone	least	out	sup	value
ain't	come	got	less	outside	sure	various
All	comes	gotten	lest	over	t	very
allow	concerning	greetings	let	overall	t's	via
allows	consequently	h	let's	own	take	viz
almost	consider	had	like	p	taken	vs
alone	considering	hadn't	liked	particular	tell	w
along	contain	happens	likely	particularly	tends	want
already	containing	hardly	little	per	th	wants
also	contains	has	look	perhaps	than	was
although	corresponding	hasn't	looking	placed	thank	wasn't
always	could	have	looks	please	thanks	way
Am	couldn't	haven't	ltd	plus	thanx	we
among	course	having	m	possible	that	we'd

[2] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop
[3] https://cran.r-project.org/web/packages/tm/index.html
[4] https://www.r-project.org/

amongst	currently	he	mainly	presumably	that's	we'll
An	D	he's	many	probably	thats	we're
And	definitely	hello	may	provides	the	we've
another	described	help	maybe	q	their	welcome
Any	despite	hence	me	que	theirs	well
anybody	did	her	mean	quite	them	went
anyhow	didn't	here	meanwhile	qv	themselves	were
anyone	different	here's	merely	r	then	weren't
anything	Do	hereafter	might	rather	thence	what
anyway	does	hereby	more	rd	there	what's
anyways	doesn't	herein	moreover	re	there's	whatever
anywhere	doing	hereupon	most	really	thereafter	when
apart	don't	hers	mostly	reasonably	thereby	whence
appear	done	herself	much	regarding	therefore	whenever
appreciate	down	hi	must	regardless	therein	where
appropriate	downwards	him	my	regards	theres	where's
Are	during	himself	myself	relatively	thereupon	whereafter
aren't	E	his	n	respectively	these	whereas
around	each	hither	name	right	they	whereby
As	edu	hopefully	namely	s	they'd	wherein
aside	Eg	how	nd	said	they'll	whereupon
Ask	eight	howbeit	near	same	they're	wherever
asking	either	however	nearly	saw	they've	whether
associated	else	i	necessary	say	think	which
At	elsewhere	i'd	need	saying	third	while
available	enough	i'll	needs	says	this	whither
away	entirely	i'm	neither	second	thorough	who
awfully	especially	i've	never	secondly	thoroughly	who's
B	Et	ie	nevertheless	see	those	whoever
Be	etc	if	new	seeing	though	whole
became	even	ignored	next	seem	three	whom
because	ever	immediate	nine	seemed	through	whose
become	every	in	no	seeming	throughout	why
becomes	everybody	inasmuch	nobody	seems	thru	will
becoming	everyone	inc	non	seen	thus	willing
been	everything	indeed	none	self	to	wish
before	everywhere	indicate	noone	selves	together	with
beforehand	Ex	indicated	nor	sensible	too	within
behind	exactly	indicates	normally	sent	took	without
being	example	inner	not	serious	toward	won't
believe	except	insofar	nothing	seriously	towards	wonder
below	F	instead	novel	seven	tried	would
beside	Far	into	now	several	tries	would
besides	few	inward	nowhere	shall	truly	wouldn't
best	fifth	is	o	she	try	x

better	first	isn't	obviously	should	trying	y
between	five	it	of	shouldn't	twice	yes
beyond	followed	it'd	off	since	two	yet
both	following	it'll	often	six	u	you
brief	follows	it's	oh	so	un	you'd
But	for	its	ok	some	under	you're
By	former	itself	okay	somebody	unfortunately	your
C	formerly	j	old	somehow	unless	yourself
c'mon	forth	just	on	someone	unlikely	Z
c's	four	k	once	something	until	zero

Table 2.1: SMART stopwords list

2.2 Graph of Words

Graph of words (GoW) is the representation of a text document as an unweighted directed graph (Rousseau and Vazirgiannis, 2013), where its nodes represent terms (words). Given a window of N successive words in a document, all terms in the window are mutually linked and each edge represents the co-occurrence of a pair of terms in the window set.

The graph is directed and each edge direction represents term order. For example, in the case of three-term window, the first term is linked with the two following terms.

In Figure 2.2 we present the GoW representation of the boxed sentence above, where text is tokenized, lowercased and a window of size $N = 3$ is adopted.

In the following chapter, we shall apply the GoW model in keyword extraction and we shall compare its performance with the keywords extracted by the BoW model, as the most frequent terms in a document.

In mathematics, and more specifically in graph theory, a graph is a structure amounting to a set of objects in which some pairs of the objects are in some sense "related"

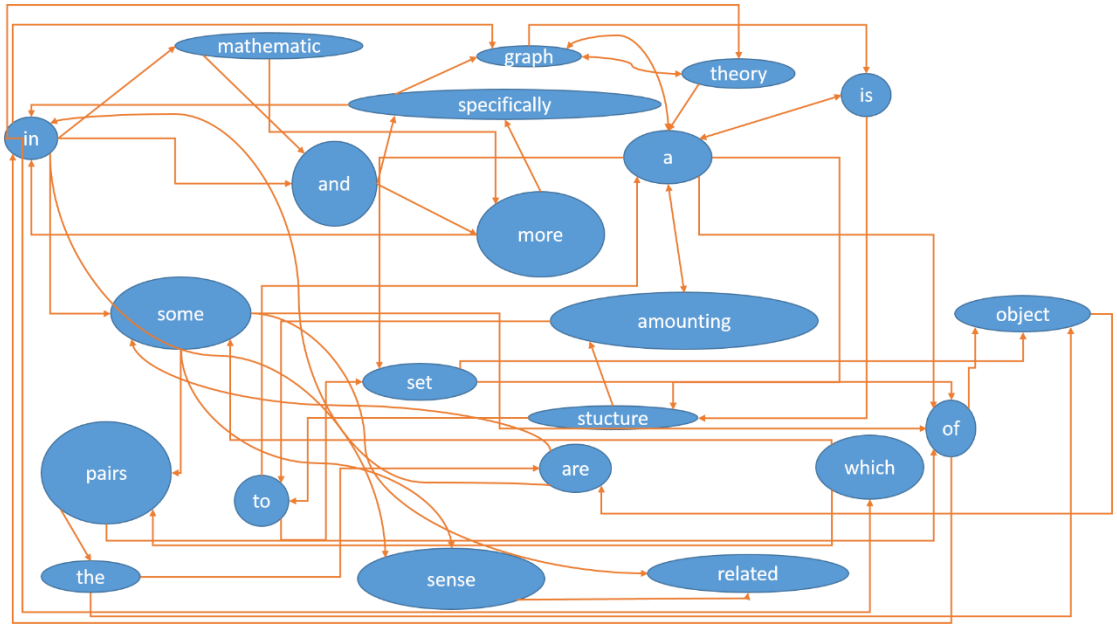


Figure 2.2: GoW representation of the graph definition from Wikipedia[5]

[5] https://en.wikipedia.org/wiki/Graph_(discrete_mathematics)

Chapter 3. Application to keyword extraction

In this chapter we apply the BoW and the GoW models in the keyword extraction problem. We compare their performance in two publicly available datasets using the evaluation measures Precision@10, mean Average Precision and Jaccard coefficient, which are introduced in Section 3.2. The comparison is done using the centrality measures and communities, presented in Chapter 1 and listed in Section 3.1. We selected these methods as the most prominent methods to identify central nodes in a graph or network. The selected evaluation measures have been used in the literature to evaluate the results of retrieval problems and Jaccard similarity is able to compare the similarity of any two sets of words. Experiments on two public datasets are presented in Section 3.4.

3.1 Methods

The methods we have selected for comparison are grouped into two main categories. Firstly, centrality measures on the formulated Graph-of-Words (GoW) and more general centrality-based scores (transitivity, coreness, high term frequency score) are employed, since they are able to rank all words in a document from the most central to the less central, according to their score in the GoW representation. Secondly, community detection algorithms on the GoW provide the largest community that contains the key nodes (words) in the GoW.

Betweenness centrality has been used in the context of keyword extraction (Beliga et al., 2014), as well as the closeness centrality (Abilhoa et al., 2014), the degree centrality (Lahiri et al., 2014), Eigenvector centrality (Boudin, 2013) and PageRank (Tsatsaronis et al., 2010). In addition, eccentricity (Xie, 2005) and coreness, transitivity (known also as clustering coefficient) and Term-Frequency (TF) scores have been examined in keyword extraction (Lahiri et al., 2014).

In this chapter, we also examine the performance of the following centrality measures in the keyword extraction problem:

- Mapping Entropy (Nie et al., 2016)
- Mapping Entropy Betweenness (MEB) (Gialampoukidis et al., 2016a)
- Mapping Entropy Closeness (MEC)

Contrary to the Edge Betweenness modularity maximization method has been used to extract communities of words (Grineva et al., 2009) Moreover, we examine the performance of the community of words, in the GoW representation, as extracted by the following approaches, which have been discussed in Chapter 1:

- Fast greedy (modularity maximization)
- Infomap (codelength minimization)
- Label Propagation
- Louvain (modularity maximization)
- Walktrap (random walks)

The results are presented in the following section.

3.2 Evaluation measures

Let $\mathcal{C}$ be the collection of documents and we denote by $\mathcal{R}$ the set of retrieved results with respect to the query q. We also denote by $\mathcal{T}$ the set of relevant documents, in terms of the annotation which is provided by the ground truth.

Definition 3.1. *Precision* is defined as the fraction of retrieved instances $\mathcal{R}$ from the collection $\mathcal{C}$ that are relevant to the query q.

Precision uses all retrieved documents, but it can also be computed at a given threshold of the top-n results returned by the system and it is called *precision at n*, usually denoted as *P@n*.

Mathematically, precision is formulated as follows:

$$precision = \frac{|relevant\ documents| \cap |retrieved\ documents|}{|retrieved\ documents|} = \frac{|\mathcal{T} \cap \mathcal{R}|}{|\mathcal{R}|} \tag{3.1}$$

where the nominator is the number of retrieved documents which are also relevant to the query and the denominator is the total number of all returned results.

Average precision

Average precision is a measure that is not set oriented and is affected by the order of relevant documents. It does not take into account only the number of retrieved documents which are relevant, but also their position in the ranking of the returned results. Average Precision (AP) is defined as:

$$AP = \frac{\sum_{n=1}^{R} P@n}{R} \tag{3.2}$$

where n is the rank of each relevant document, R is the total number of relevant documents, and $P@n$ is the precision of the top-n retrieved documents.

mean Average Precision

Given a set of Q queries, mean Average Precision (mAP) is defined as the mean of all Average Precision scores for each query:

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{3.3}$$

where, $AP(q)$ is the Average Precision for the query q.

Definition 3.2. *Recall* is defined as the fraction of relevant instances that are retrieved $|\mathcal{T} \cap \mathcal{R}|$ from the collection $\mathcal{C}$ to the total number of relevant documents $\mathcal{T}$:

$$recall = \frac{|relevant\ documents| \cap |retrieved\ documents|}{|relevant\ documents|} = \frac{|\mathcal{T} \cap \mathcal{R}|}{|\mathcal{T}|} \tag{3.4}$$

It is trivial to achieve recall of 100% by returning all documents in response to any query, where all relevant documents are retrieved:

$$recall = \frac{|\mathcal{T} \cap \mathcal{R}|}{|\mathcal{T}|} = \frac{|\mathcal{T}|}{|\mathcal{T}|} = 1 \tag{3.5}$$

Therefore, recall alone is not enough and one needs to combine precision and recall, as done, for example, in the measure of F_1 score.

Definition 3.3. F_1*-score* is defined as a measure that combines precision and recall as follows:

$$F_1 = 2\frac{precision * recall}{precision + recall} \tag{3.6}$$

F_1 score is approximately the average of precision and recall when they are close, and is more generally the harmonic mean.

Jaccard similarity

The Jaccard index, also known as the Jaccard similarity coefficient, is a statistic used for comparing the similarity of two sample sets and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.7}$$

3.3 Dataset Description

The datasets we have selected for comparison are, firstly, the Fao780[6] dataset which contains 779 documents and the CiteULike180[7] dataset with 183 text documents, tagged by 152 taggers. The CiteULike dataset has 183 publications crawled from CiteULike, and keywords assigned by different CiteULike users who saved these publications. The other dataset, FAO780, has 779 FAO publications with Agrovoc terms from official documents of the Food and Agriculture Organization of the United Nations (FAO).

3.4 Settings

Firstly, we remove punctuation and we transform all letters to lowercase. Numbers are also removed, as well as the English stopwords, which are common words that are repeated (e.g. “the”, “a”, “and”) without adding meaning to the document, as already presented in Section 2.1 (SMART stopwords list). Moreover, we stem each word, i.e. we remove the ending of the word, so as to keep only the word’s stem. Afterwards, we construct the graph of words, which has as nodes the words of our document. Two nodes take link if a word follows the other, i.e. any two terms of a bi-gram (N=2) are connected. We also examine the performance of the keyword extraction problem, by linking the terms of tri-grams (N=3).

In all datasets, we keep the top-20 keywords for each selected centrality score (Betweenness, Closeness, Degree, Eigenvector, Page Rank, Mapping Entropy, MEB, MEC, Coreness, Transitivity, Eccentricity) and for the top-20 most frequent terms (TF scores). In the case of the most informative community of the constructed graph of words, we use five prominent community detection algorithms (Fast greedy, Infomap, Label Prop, Louvain and Walktrap).

[6] https://github.com/zelandiya/keyword-extraction-datasets
[7] https://github.com/snkim/AutomaticKeyphraseExtraction

3.5 Results

FAO documents have more unstructured text than CiteULike documents, as shown in Appendix C, where we present two sample text documents from these datasets. In the case of structured text (CiteULike), we observe that the GoW representation performs better than the simple statistical term frequency scores. On the other hand, in the FAO dataset, TF scores count the most frequent words and are able to identify the most critical words in each document. In structured text, the order of words is very important because links are added between a word and its N successive words. Hence, the GoW is superior to the Bag of Words representation in the case of structured sentences.

Given the GoW representation, we observe that when N=3 the results are better than N=2, where N is the number of successive words that are linked to any word. However, the linking of more words than N=3 successive words, makes the graph of words almost complete, so centralities become identical and the graph has only one community (all the graph).

Among the centrality measures, closeness centrality performs better than the other measures. In the case of N=2, Mapping Entropy Betweenness centrality has larger Jaccard index than all other methods.

Among the community detection approaches, the Infomap communities contain the most important words on average and therefore obtain higher Jaccard, Average Precision and P@10.

Community detection approaches are not superior to centrality scores, in all cases examined.

Our proposed Mapping Entropy Closeness (MEC) centrality measure is the second most performing keyword extraction approach, in the case of Jaccard index, following the Mapping Entropy Betweenness (MEB) scores.

N=2	Citeulike180			Fao780		
Method	**Jaccard**	**Average Precision**	**P@10**	**Jaccard**	**Average Precision**	**P@10**
Betweenness	0.1531 ± 0.0598	0.3795 ± 0.1401	0.3486 ± 0.1398	0.1619 ± 0.0734	0.3459 ± 0.1500	0.3112 ± 0.1473
Closeness	0.1531 ± 0.0622	$\mathbf{0.3890} \pm 0.1425$	$\mathbf{0.3552} \pm 0.1413$	0.1656 ± 0.0781	0.3565 ± 0.1547	0.3212 ± 0.1540
Degree	0.1566 ± 0.0611	0.3842 ± 0.1390	0.3492 ± 0.1410	0.1671 ± 0.0777	0.3533 ± 0.1538	0.3208 ± 0.1508
Eigenvector	0.1446 ± 0.0659	0.3606 ± 0.1453	0.3525 ± 0.1421	0.1649 ± 0.0792	0.3526 ± 0.1570	0.3158 ± 0.1549
Page Rank	0.0508 ± 0.0313	0.3831 ± 0.1399	0.3492 ± 0.1410	0.1669 ± 0.0772	0.3488 ± 0.1530	0.3173 ± 0.1503
Mapping Entropy	0.1557 ± 0.0613	0.3821 ± 0.1394	0.3519 ± 0.1406	0.1669 ± 0.0780	0.3515 ± 0.1533	0.3191 ± 0.1502
MEB	$\mathbf{0.1598} \pm 0.0625$	0.3860 ± 0.1378	0.3530 ± 0.1354	0.0674 ± 0.0451	0.1762 ± 0.1180	0.1469 ± 0.1009
MEC	0.1567 ± 0.0622	0.3839 ± 0.1389	0.3503 ± 0.1402	0.0678 ± 0.0460	0.1753 ± 0.1178	0.1477 ± 0.1009
Coreness	0.1098 ± 0.5110	0.2857 ± 0.1364	0.3508 ± 0.1568	0.0839 ± 0.0487	0.1802 ± 0.0994	0.2855 ± 0.1556
Transitivity	0.0000 ± 0.0000	0.0182 ± 0.0469	0.0164 ± 0.0426	0.0067 ± 0.0154	0.0221 ± 0.0559	0.0171 ± 0.0422
Eccentricity	0.0015 ± 0.0062	0.0026 ± 0.0157	0.0027 ± 0.0163	0.0003 ± 0.0033	0.0004 ± 0.0054	0.0004 ± 0.0062
TF score	0.1613 ± 0.0648	0.3877 ± 0.1421	0.3530 ± 0.1386	$\mathbf{0.1781} \pm 0.0843$	$\mathbf{0.3725} \pm 0.1603$	$\mathbf{0.3392} \pm 0.1614$
Fast greedy	0.0215 ± 0.0164	0.0649 ± 0.0500	0.1656 ± 0.1459	0.0100 ± 0.0116	0.0297 ± 0.0303	0.1163 ± 0.1114
Infomap	0.0402 ± 0.0248	0.1258 ± 0.0762	0.2749 ± 0.1770	0.0205 ± 0.0220	0.0586 ± 0.0581	0.2258 ± 0.1462
Label Prop	0.0158 ± 0.0088	0.0411 ± 0.0203	0.2754 ± 0.1693	0.0074 ± 0.0069	0.0219 ± 0.0153	0.2100 ± 0.1420
Louvain	0.0193 ± 0.0167	0.0600 ± 0.0538	01421 ± 0.1415	0.0107 ± 0.0130	0.0320 ± 0.0359	0.0992 ± 0.1054
Walktrap	0.0332 ± 0.0171	0.0941 ± 0.0459	0.3060 ± 0.1846	0.0176 ± 0.0173	0.0504 ± 0.0412	0.2144 ± 0.1439

Table 3.1: Jaccard, Average Precision and P@10 results for linking N=2 successive words

N=3	Citeulike180			Fao780		
Method	**Jaccard**	**Average Precision**	**P@10**	**Jaccard**	**Average Precision**	**P@10**
Betweenness	0.1609 ± 0.0633	0.3854 ± 0.1431	0.3519 ± 0.1441	0.1671 ± 0.0748	0.3568 ± 0.1505	0.3213 ± 0.1504
Closeness	$\mathbf{0.1658} \pm 0.0617$	$\mathbf{0.4034} \pm 0.1447$	$\mathbf{0.3776} \pm 0.1490$	0.1731 ± 0.0819	0.3678 ± 0.1560	0.3326 ± 0.1558
Degree	0.1648 ± 0.0621	0.3993 ± 0.1406	0.3661 ± 0.1404	0.1744 ± 0.0806	0.3671 ± 0.1543	0.3304 ± 0.1532
Eigenvector	0.1542 ± 0.0629	0.3791 ± 0.1445	0.3448 ± 0.1428	0.1711 ± 0.0818	0.3662 ± 01589	0.3291 ± 0.1590
Page Rank	0.1645 ± 0.0662	0.3982 ± 0.1401	0.3678 ± 0.1395	0.1740 ± 0.0807	0.3641 ± 0.1542	0.3286 ± 0.1530
Mapping Entropy	0.1644 ± 0.0632	0.3974 ± 0.1404	0.3650 ± 0.1394	0.1746 ± 0.0807	0.3662 ± 0.1544	0.3295 ± 0.1540
MEB	0.1638 ± 0.0619	0.3963 ± 0.1397	0.3661 ± 0.1435	0.1723 ± 0.0776	0.3627 ± 0.1527	0.3293 ± 0.1530
MEC	0.1648 ± 0.0636	0.3886 ± 0.1407	0.3683 ± 0.1402	0.1745 ± 0.0803	0.3671 ± 0.1544	0.3295 ± 0.1527
Coreness	0.1066 ± 0.0481	0.2637 ± 0.1208	0.3694 ± 0.1682	0.075 ± 0.0440	0.1595 ± 0.0848	0.2796 ± 0.1542
Transitivity	0.0015 ± 0.0062	0.0025 ± 0.0161	0.0022 ± 0.0147	0.0001 ± 0.0050	0.0015 ± 0.0130	0.0014 ± 0.0118
Eccentricity	0.0016 ± 0.0067	0.0022 ± 0.0124	0.0033 ± 0.0179	0.0006 ± 0.0045	0.0010 ± 0.0090	0.0006 ± 0.0080
TF score	0.1613 ± 0.0648	0.2637 ± 0.1208	0.3530 ± 0.1386	$\mathbf{0.1781} \pm 0.0843$	$\mathbf{0.3725} \pm 0.1603$	$\mathbf{0.3392} \pm 0.1614$
Fast greedy	0.0196 ± 0.0146	0.0565 ± 0.0399	0.1792 ± 0.1475	0.0086 ± 0.0098	0.0255 ± 0.0257	0.1167 ± 0.1169
Infomap	0.0283 ± 0.0167	0.0865 ± 0.0490	0.2995 ± 0.1903	0.014 ± 0.0145	0.0407 ± 0.0393	0.2248 ± 0.1423
Label Prop	0.0151 ± 0.0077	0.0394 ± 0.0181	0.2689 ± 0.1696	0.0072 ± 0.0066	0.0216 ± 0.0147	0.2089 ± 0.1412
Louvain	0.0160 ± 0.0154	0.0464 ± 0.0444	0.1235 ± 0.1294	0.0098 ± 0.0111	0.0288 ± 0.0298	0.1141 ± 0.1166
Walktrap	0.0280 ± 0.0166	0.0809 ± 0.0436	0.2891 ± 0.1895	0.0140 ± 0.0136	0.0414 ± 0.0347	0.1979 ± 0.1418

Table 3.2: Jaccard, Average Precision and P@10 results for linking N=3 successive words

Chapter 4. Text clustering

In this chapter we present popular clustering approaches that have been used in the context of text clustering. Firstly, a density-based algorithm called DBSCAN is reported and secondly, k-means clustering is presented. Hierarchical clustering is thirdly presented and finally we refer to Latent Dirichlet Allocation as a well performing method in topic modeling. Evaluation measures in clustering are also described, such as Normalized Mutual Information, Variation of Information, Rand and adjusted Rand indices. Finally, we examine which of the clustering approaches perform better in a public dataset of news articles, as well as in a collection of Twitter posts in the short text clustering problem.

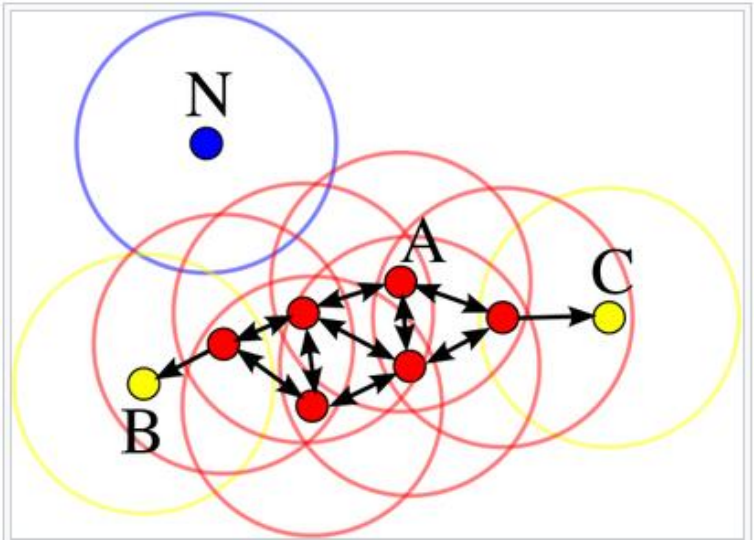
4.1 Methods

Density-based clustering

The DBSCAN algorithm is one of the first approaches in density-based clustering (Ester et al., 1996). It is based on the definition of core, border and noise points, which are defined using the notion of reachability. Density-reachable and directed density-reachable points (items) are defined with respect to the parameters ε and $MinPts$.

Definition 4.1: (ε-neighborhood of a point) The ε-neighborhood of a point p, denoted by $N_\varepsilon(p)$, is defined by $N_\varepsilon(p) = \{q \in D | dist(p,q) \leq \varepsilon\}$ and D is the set of all points.

Instead of requiring for each point in a cluster having at least a minimum number ($MinPts$) of points in an ε-neighborhood of that point, border points also appear. There are two kinds of points in a cluster, points inside of the cluster (core points) and points on the border of the cluster (border points). The difference between core and border points is visualized in Figure 4.1.



Figure[8] 4.1: Border, core and noise points in DBSCAN.

An ε-neighborhood of a border point contains significantly less points than an ε-neighborhood of a core point. Hence, DBSCAN would have to set the minimum number of points to a relatively low value in order to include all points belonging to the same cluster. DBSCAN requires that for every point p in a cluster C there is a point q in C, so that p is inside of the ε-neighborhood of q and $N_\varepsilon(q)$, contains at least $MinPts$ points. This definition is elaborated in the following.

Definition 4.2: (directly density-reachable) A point p is directly density-reachable from a point q with respect to ε, $MinPts$ if

1) $p \in N_\varepsilon(q)$

2) $|N_\varepsilon(q)| \geq \mathrm{MinPts}$ (core point condition).

[8] https://en.wikipedia.org/wiki/DBSCAN

Obviously, directly density-reachable is symmetric of or pairs of core points. In general, however, it is not symmetric if one core point and one border point are involved.

Definition 4.3: (density-reachable) A point p is density reachable from a point q with respect to ε and $MinPts$ if there is a chain of points $p_l \ldots . p_n, p_l = q, p_n = p$ such that p_{i+l} is directly density-reachable from p_l.

Density-reachability is a canonical extension of direct density-reachability. This relation is transitive, but it is not symmetric. Although not symmetric in general, it is obvious that density-reachability is symmetric for core points.

The lack of symmetry in both border and core points requires the definition of density-connectivity, so that any two border points in the cluster are related. The motivation of the following definition comes from the fact that two border points of the same cluster C are possibly not density reachable from each other. However, there must be a core point in C from which both border points of C are density-reachable.

Definition 4.4: (density-connected) A point p is density connected to a point q with respect to ε and $MinPts$ if there is a point o such that both, p and q are density-reachable from o with respect to ε and $MinPts$.

Density-connectivity is a symmetric relation. Using Definitions 4.1, 4.2, 4.3, 4.4 the density-based notion of a cluster is defined to be a set of density connected points which is maximal with respect to density-reachability:

Definition 4.5: (cluster) Let D be a collection of points in any n-dimensional space. A cluster C with respect to ε and $MinPts$ is a non-empty subset of D satisfying the following conditions:
1) $\forall p, q: if\ p \in C$ and q is density-reachable from p with respect to ε, $MinPts$, then $q \in C$.
2) $\forall p, q \in C: p$ is density-connected to q with respect to ε and $MinPts$.

Noise is defined relative to a given set of clusters, as the set of points in D which does not belong to any of its clusters.

Definition 4.6: (noise) Let $C_1, \ldots, C_k$ be the clusters of the database D with respect to parameters ε_i and $MinPts_i, i = 1 \ldots . k$. Then we define the noise as the set of points in the database D not belonging to any cluster C_i, i.e. noise = $\{p \in D | \forall\, i: p \notin C_i\}$.

Remark: A cluster C with respect to ε and $MinPts$ contains at least $MinPts$ points.

The cluster is non-empty by definition and therefore it has at least one point p, that is density-connected to itself via some point o. Thus, o has to satisfy the core point condition, i.e. there is an ε-neighborhood of o that contains at least $MinPts$ points.

The DBSCAN algorithm discovers a cluster in a two-step approach given the parameters ε and $MinPts$. Firstly, DBSCAN chooses an arbitrary point from the collection of points, satisfying the core point condition, as a seed. Secondly, DBSCAN retrieves all points that are density-reachable from the seed, obtaining the cluster containing the seed. These two steps are repeated until all points are either assigned to a cluster or marked as noise.

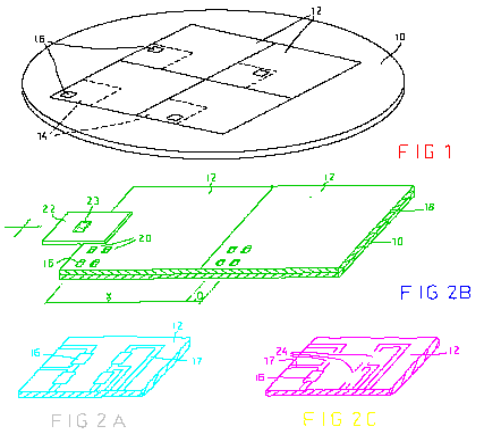


Figure 4.2: An example of a patent image and its segmentation into clusters by DBSCAN using one color per cluster.

Hierarchical clustering

Strategies for hierarchical clustering are two:

- **Agglomerative** is a "bottom up" approach. Each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive** is a "top down" approach. All observations start in one cluster, and clusters are split recursively as one moves down the hierarchy.

In both cases, the results are presented in a dendrogram. More specifically, agglomerative hierarchical clustering is performed as follows. Given a set of N items to be clustered and $N \times N$ distance matrix, the process of Johnson's (1967) hierarchical clustering in this:

1. Start by assigning each item in its own cluster, so that if you have N items, you have N clusters each containing just one item.
2. Find the distance for any pair of items they contain
3. Find the closest pair of clusters and merge them into a single cluster, so that you have one less cluster
4. Compute distances between the new cluster and each of the old clusters
5. Repeat 3+4 until all items are clustered into a single cluster of size N

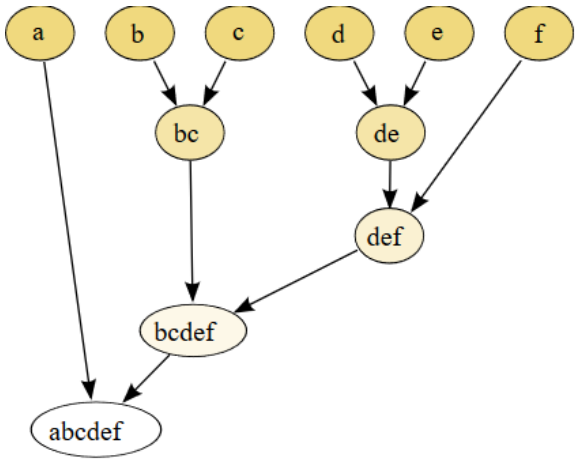


Figure 4.3: Hierarchical agglomerative clustering example[9]

[9] https://upload.wikimedia.org/wikipedia/commons/a/ad/Hierarchical_clustering_simple_diagram.svg

Alternative to the agglomerative hierarchical clustering is the divisive approach. In this method we initially assign all of the observations to a single cluster and we proceed recursively by splitting each cluster, until there is one cluster for each observation.

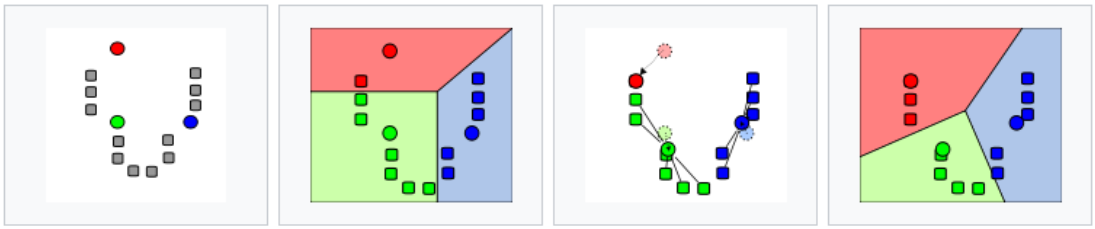
K-means clustering

k-means clustering groups points on a n-dimensional space into k clusters. The purpose of k-means algorithm is to minimize the overall variance within each group, measured by the square error function:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} |x_j - \mu_i|^2 \tag{4.1}$$

Where there are k clusters $S_i, i = 1,2, \dots, k$ and μ is the central point.

1. Set the number of clusters
2. Random creation k-clusters and choose of central points of the clusters
3. Transfer each point to the central of the nearest cluster
4. Calculation of new central points
5. Repeat until it converges



Figure[10] 4.4: Clustering 12 points into 3 clusters by k-means clustering

The quality of the solution depends on the initial set of clusters, i.e. the initial selection of the k-centers. The algorithm tends to converge quickly but the number of clusters should be defined from the beginning of the algorithm.

[10] https://en.wikipedia.org/wiki/K-means_clustering

In R, the algorithm of Hartigan and Wong (1979) is the default algorithm. K-means is usually referred to a specific algorithm rather than the general method (MacQueen, 1967) but sometimes that given by Lloyd (1957) and Forgy (1965). Trying several random starts is often recommended, due to the randomness of the initial choice of centers.

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic topic model, which has been introduced in the context of text clustering (Blei et al., 2003). It is assumed that documents are represented using the Bag-of-Words (unigrams, n-grams, etc.) model and that each topic is a distribution over terms (words) in a fixed vocabulary. Each topic contains different words with different probabilities and each topic is characterized by a distribution over words. Note that the order of words does not matter and that "topic" is a distribution over terms. The word frequencies are observed variables, which are used to estimate topic distributions based on the statement that Dirichlet[11] distribution is the conjugate prior[12] distribution of the multinomial distribution, defined as follows:

$$f(x_1, x_2, \dots, x_k; n; p_1, p_2, \dots, p_k) = \frac{n!}{x_1!\, x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \tag{4.2}$$

where

$$\sum_{i=1}^{k} x_i = n \quad and \quad \sum_{i=1}^{k} p_i = 1 \tag{4.3}$$

LDA has been generalized to nonparametric Bayesian approaches, such as the hierarchical Dirichlet process (Teh et al., 2004) and DP-means (Kulis and Jordan, 2011), which predict the number of topics. The extraction of the correct number of topics is equivalent to the estimation of the correct number of clusters in a dataset and the majority vote among 30 clustering indices has recently been proposed in (Charrad et al., 2014) as an indicator for the

[11] https://en.wikipedia.org/wiki/Dirichlet_distribution
[12] https://en.wikipedia.org/wiki/Conjugate_prior

number of clusters in a dataset. Recently, a density-based clustering approach has been introduced, namely DBSCAN-Martingale (Gialampoukidis et al., 2016d), which clusters bi-grams to first estimate the number of topics and then assign text documents to topics using Latent Dirichlet Allocation on the set of all extracted unigrams.

4.2 Evaluation measures

For the evaluation of clustering the most popular measures are the Normalized Mutual Information, Rand Index, Adjusted Rand Index and Variation of Information. Let us suppose two partitions of the dataset $C = (C_1, C_2, \dots, C_k)$ and $C' = (C'_1, C'_2, \dots, C'_{k'})$ and their corresponding number of clusters k and k'. Let also N be the number of items – points, and n_k, $n'_{k'}$ the number of items which are both of two clusters C_i and C'_i. Evaluation measures belong in two main categories; pair counting measures and evaluation measures based on information theory.

Pair counting measures involve Rand Index and adjusted Rand Index.

Rand index

Measures which based on pair counting depends on number of pairs of items, which are at the same cluster in both of two clusters. Specifically N_{11} is defined as the number of pairs of items which are at the same cluster in both partitions, where $N_{10}(N_{01})$ is the number of pairs of nodes which are in the same cluster at partition $C(C')$ and at different cluster in partition $C'(C)$, and N_{00} is the number of pairs of items which are at different clusters at both partitions. In total:

$$N_{11} + N_{00} + N_{10} + N_{01} = n(n-1)/2 \tag{4.4}$$

Rand (1971) defined Rand Index measure as the ratio of the number of pairs of items which are on both partitions to the total number of pairs of items:

$$R(C, C') = \frac{N_{11} + N_{00}}{N_{11} + N_{00} + N_{10} + N_{01}} = \frac{2(N_{11} + N_{00})}{n(n-1)} \tag{4.5}$$

Adjusted Rand Index

Hubert and Arabie (1985) define the Adjusted Rand Index:

$$AR(C,C') = \frac{R(C,C') - E[R]}{1 - E[R]} \tag{4.6}$$

$$= \frac{\sum_{k=1}^{K}\sum_{k'=1}^{K'}\binom{n_{kk'}}{2} - [\sum_{k=1}^{K}\binom{n_k}{2}]\,[\sum_{k'=1}^{K'}\binom{n_{k'}}{2}]/\binom{n}{2}}{[\sum_{k=1}^{K}\binom{n_k}{2} + \sum_{k'=1}^{K'}\binom{n_{k'}}{2}]/2 - [\sum_{k=1}^{K}\binom{n_k}{2}][\sum_{k'=1}^{K'}\binom{n_{k'}}{2}]/\binom{n}{2}}$$

Where $E[R]$ is how similar are which expected due to luck. There are three cases:

- $AR = 0$ if clusters are independed
- $AR = 1$ if clusters are same
- $AR < 1$ if there is strong deviation

Measures which based on information theory have as central idea the fact that if two partitions are very close, then the necessary information amount for one partition given the other is small.

Mutual information

If we choose random an item from a cluster, how much is the uncertainty about the cluster that it belongs to? Assuming that each node has the same probability to be chosen, then the probability of a chosen item to be in the cluster C_k is:

$$P(k) = \frac{n_k}{n} \tag{4.7}$$

Hence, there is a discrete variable which take values K and linked to cluster C. The uncertainty is equal to the Shannon entropy of this variable:

$$H(C) = -\sum_{k=1}^{K} P(k) log P(k) \tag{4.8}$$

Entropy takes always negative values. Takes 0 only in case of there is not uncertainty, ie in case of the cluster C has only one community. Entropy counts in bits, if the uncertainty is 1

bit $K = 2$ and $P(1) = P(2) = 0.5$. Uncertainty is not depended on the number of items, but it depends on rate of items which belongs in each cluster.

Mutual information between two partitions C and C' is given by the formula:

$$I(C, C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') log \frac{P(k, k')}{P(k)P'(k')} \tag{4.9}$$

where $P(k)$ is the probability distribution of random variable which linked to cluster C and $(k, k') = |C_k \cap C_{k'}|/n$, the common probability distribution of random variables of clusters.

Mutual information is the information for a partition given the information for the other. If we know the uncertainty $H(C')$, with which an item belongs to a cluster in cluster C', and if we also know in which cluster belongs the same item in cluster C, then the increase of uncertainty in cluster C' is equal to $I(C, C')$.

Mutual information between two clusters C and C' is

$$\begin{aligned} I(C, C') &= H(C) - H(C|C') = H(C') - H(C'|C) = H(C) + H(C') - H(C, C') \\ &= H(C, C') - H(C|C') - H(C', C) \end{aligned} \tag{4.10}$$

where $H(C)$ and $H(C')$ are entropies of C and C', $H(C, C') = H(C) + H(C'|C) = H(C') + H(C|C')$ is the jointly information, $H(C|C')$ and $H(C'|C)$ are the conditional entropies of C and C'. Mutual information of two random variables is always no negative and symmetrical

$$I(C, C') = I(C', C) \geq 0 \tag{4.11}$$

Moreover, mutual information cannot be larger than entropy's value:

$$I(C, C') \leq \min(H(C), H(C')) \tag{4.12}$$

Normalized Mutual Information

Mutual information, as a similarity measure between two cluster structures is not suitable as a universal evaluation measure. Danon et al., (2005) proposed the normalized mutual information as a similarity measure of two partitions. Based on the construction of a confusion matrix A, in which rows corresponded to known clusters and columns to clusters which coming from a partitioning (clustering) algorithm. Each A_{ij} of the matrix A corresponds to the number of items of known cluster i which there are in detected in cluster j. From this matrix we compute normalized mutual information as follows:

$$NMI(C, C') = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} A_{ij}\log(\frac{A_{ij}A}{N_{i.}N_{.j}})}{\sum_{i=1}^{C_A} N_{i.}\log\left(\frac{A_{i.}}{A}\right) + \sum_{i=1}^{C_B} A_{.j}\log\left(\frac{A_{.j}}{A}\right)} \quad (4.13)$$

where C_A and C_B is the number of known detected communities, $A_{i.}$ Is the sum of elements of i row, $A_{.j}$ is the sum of elements of j column. Normalized mutual information, gets maximum value $NMI(C, C') = 1$, in case that the two partitions are exactly the same and gets the value $NMI(C, C') = 0$ in case of the two partitions C, C' are completely different.

Variation of Information

Variation of information (VI) adopted from (Meila 2003) as a similarity measure of two clusters and it is defined as:

$$VI(C, C') = H(C) + H(C') - 2I(C, C') \quad (4.14)$$

It could also be written as:

$$VI(C, C') = [H(C) - I(C, C')] + [H(C') - I(C, C')] \quad (4.15)$$

Two of summary terms correspond to under condition entropies $H(C\backslash C')$ and $H(C'\backslash C)$. The first term give us the information amount which we loose on cluster C, going from C to C', and the other term give us the information amount which we win on C', going from C to C'.

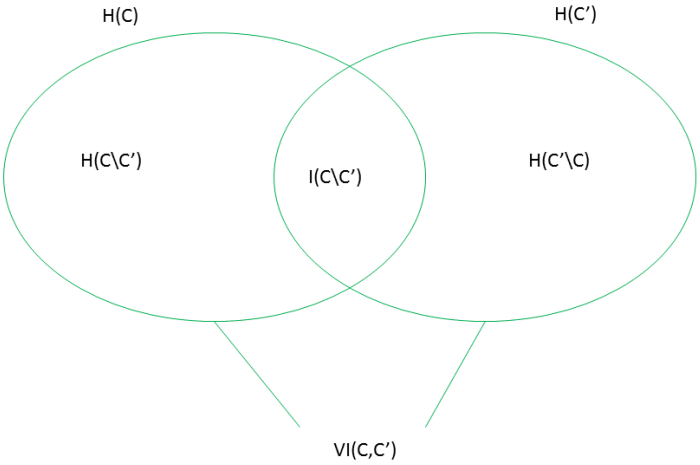


Figure 4.5: Venn diagram for the demonstration of Variation of Information (VI)

Moreover, it holds that $VI(C,C') = H(C|C') + H(C'|C)$, where $H(C|C') = H(C) - I(C,C')$ and $H(C'|C) = H(C') - I(C,C')$ the under condition entropies of two clusters. Variation of information is non negative and symmetric $VI(C,C') = VI(C',C) \geq 0$ while it is $VI(C,C') = 0$ if only two partitions are the same. Variation of information gets the maximum value $I(C,C') = logn$, when giving partition C', we do not get any information for cluster C and opposite, giving partition C, we do not get any information for the partition C' , i.e. when:

$$I(C,C') = 0 \tag{4.16}$$

VI's value does not depend on the number of items, but on the percentage of items which belong to each cluster. Variation of information sums the amount of information which is needed in order to descript partition C, when it is known partition C' and its information amount which is needed in order to describe the partition C' when C is known. Its maximum

value is $VI(C,C') = logn$, where n is the number of nodes of network. To compare different size networks is possible to normalize VI with $logn$.

4.3 Application in clustering news articles into topics

We compare the above methods with the community detection of the GoW representation.

Dataset description

The dataset we use for comparison are downloaded from online resources[13]. The WikiRef220 dataset contains 220 news articles, which are references to specific Wikipedia pages. The selected topics of the WikiRef220 dataset (and the number of articles per topic) are Paris Attacks November 2015 (36), Barack Obama (5), Premier League (37), Cypriot Financial Crisis 2012-2013 (5), Rolling Stones (1), Debt Crisis in Greece (5), Samsung Galaxy S5 (35), Greek Elections June 2012 (5), smartphone (5), Malaysia Airlines Flight 370 (39), Stephen Hawking (1), Michelle Obama (38), Tohoku earthquake and tsunami (5), NBA draft (1), U2 (1), Wall Street (1). The topics Barack Obama, Cypriot Financial Crisis 2012-2013, Rolling Stones, Debt Crisis in Greece, Greek Elections June 2012, smartphone, Stephen Hawking, Tohoku earthquake and tsunami, NBA draft, U2 and Wall Street appear no more than 5 times and therefore, they are regarded as noise. The remaining 5 topics of WikiRef220 are:

- Paris Attacks November 2015
- Premier League
- Malaysia Airlines Flight 370
- Samsung Galaxy S5
- Michelle Obama

[13] https://www.multisensorproject.eu/achievements/datasets/

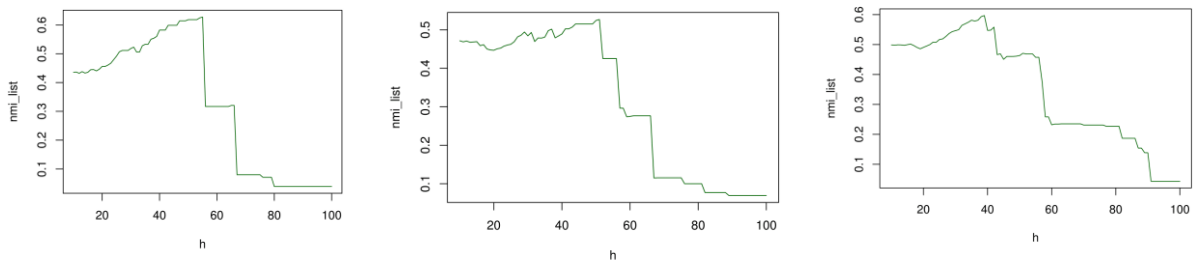
The WikiRef186 dataset (4 topics) is the WikiRef220 without 34 documents related to "Malaysia Airlines Flight 370" and the WikiRef150 dataset (3 topics) is the WikiRef186 without the 36 documents related to "Paris Attacks".

Settings

In these three datasets we apply five most popular clustering methods (DBSCAN, k-means, hierarchical clustering, LDA, graph-based clustering) and then we evaluate them with four indices, namely NMI, VI, Rand and Adjusted Rand, which have been presented in Section 4.2. We remove punctuation, numbers and stopwords, we transform all letters to lowercase, and we stem each word.

In k-means clustering and in LDA, the number of clusters is determined from the number of clusters in each dataset, known a priori from its ground-truth annotation. In DBSCAN clustering, we adopt the parameters of $MinPts$ and ε which are obtained from (Gialampoukidis et al., 2016d), and we keep the best performance per density level ε.

The best performance for the hierarchical clustering is achieved at the height (h) in which the maximum NMI score is attained, as demonstrated in Figure 4.6.

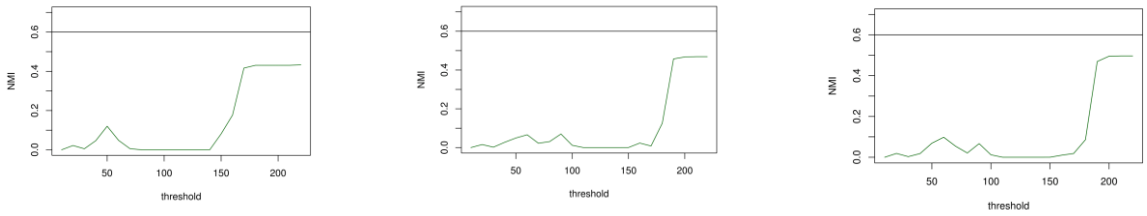


(a) wikiref150 (b) wikiref186 (c) wikiref220

Figure 4.6: NMI diagrams for each height h of the dendrogram cut.

In Figure 4.7 we present an alternative clustering approach based on a graph of items, combined with a community detection algorithm (modularity maximization) that clusters all items into communities. The graph is formulated from the mutual (Euclidian) distances of all items using a threshold distance to link or not link any two items (nodes). This clustering

approach shall be called Louvain, since the adopted community detection algorithm is based on the maximization of modularity. The best performance for the Louvain clustering is achieved at the threshold in which the maximum NMI score is attained, as demonstrated in Figure 4.7.



(a) wikiref150 (b) wikiref186 (c) wikiref220

Figure 4.7: NMI diagrams for each distance threshold.

Results

In Table 4.1 we present the evaluation results for all datasets examined. In the WikiRef150 dataset the best performing clustering method is hierarchical clustering, as indicated by all evaluation measures. In the other two datasets LDA method is by far the best method in all evaluation measures apart from Rand index. This could happen because of the nature of the dataset WikiRef150 with only three clusters.

Regarding the evaluation by the Rand index, we observe that its highest values appear in hierarchical clustering in all cases examined. Since hierarchical clustering joins pairs of very close items it is expected that any random pair of items will have both members in the same cluster. Taking into account that Rand index is a pair counting measure (examines whether two members of random pair of items belong to the same cluster or in different ones), Rand index could be biased due to the relation between its definition and the hierarchical clustering procedure.

Although hierarchical clustering has shown better performance than LDA and other methods in the WikiRef150 dataset, it is hard to estimate the optimal height h in datasets with no ground-truth annotation.

	WikiRef150 (3-clusters)				WikiRef186 (4-clusters)				WikiRef220 (5-clusters)			
Method	**NMI**	**VI**	**Rand**	**Adj_rand**	**NMI**	**VI**	**Rand**	**Adj_rand**	**NMI**	**VI**	**Rand**	**Adj_rand**
DBSCAN	0.0431	2.0281	0.5043	0.0109	0.0431	2.1460	0.4491	0.0072	0.0431	2.2853	0.4223	0.0092
Hierarchical	**0.6280**	**0.9687**	**0.7870**	**0.4763**	0.5270	1.5833	**0.8051**	0.2660	0.5969	1.7466	**0.8390**	0.2736
K-means	0.3047	1.3312	0.4834	0.1334	0.2209	1.5816	0.3762	0.0619	0.1929	1.7488	0.3256	0.0422
LDA	0.5611	1.0583	0.7643	0.4628	**0.6598**	**0.9091**	0.7984	**0.5086**	**0.6052**	**1.1011**	0.7350	**0.3873**
Louvain	0.4331	1.3851	0.7545	0.0736	0.4689	1.6084	0.8038	0.0252	0.4972	1.7909	0.8368	0.0526

Table 4.1: Evaluation results for clustering news articles

4.4 Application in clustering Twitter posts

We compare LDA text clustering, which has shown great performance in the text clustering problem (Section 4.3), with graph-based representations of the collection using the Jaccard similarity to construct the graph of documents, followed by a community detection approach (Louvain method for modularity maximization) to cluster the documents.

Dataset description

We collected 500 social media posts from Twitter using the library twitteR[14] in R. The selected topics are "Trump", "PippasWedding", "examseason", "SpecialOlympics" and "earthquake", where we kept 100 posts per topic in our analysis. The topics were selected as five of the most popular trending topics with posts in English language and were crawled in May 21st, 2017.

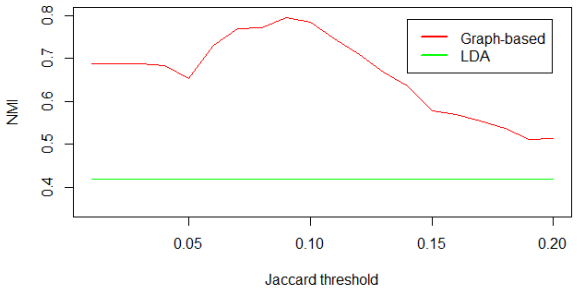
[14] https://cran.r-project.org/web/packages/twitteR/index.html

Settings

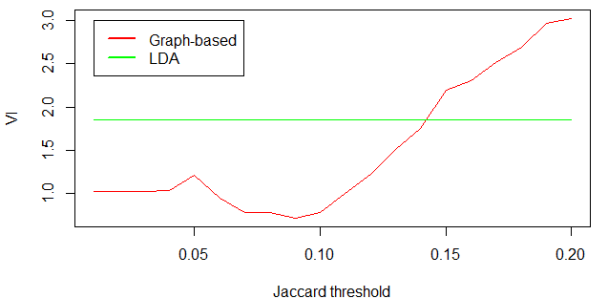
Firstly, we remove punctuation and we transform all letters to lowercase. In addition, we remove numbers and stopwords (Section 2.1) and we stem each word. Afterwards, we construct the graph which has as nodes each document (post). Documents take link according to the Jaccard similarity as follows: if two documents have Jaccard similarity larger than a considered threshold then they are linked. We grid threshold values in $\{0.01, 0.02, \dots, 0.20\}$ and present each evaluation measure per all threshold values. For the community detection approach on the constructed graph, we adopt the popular modularity maximization algorithm Louvain (Section 1.3).

Results

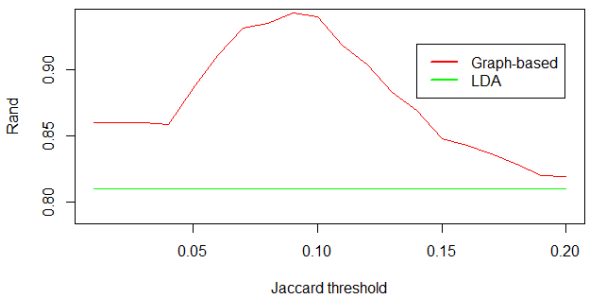
The comparison between LDA and the graph-based clustering approach in the collected Twitter posts is presented in Figure 4.8.



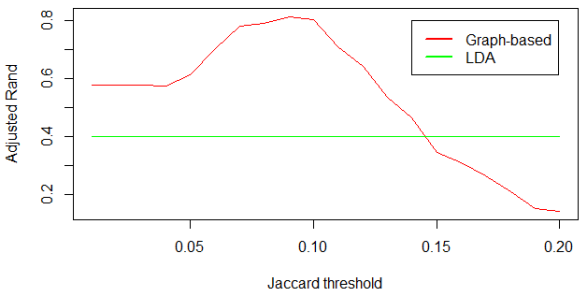
(a) NMI scores



(b) VI scores



(c) Rand scores



(d) Adjusted Rand scores

Figure 4.8: Comparison between LDA and graph-based clustering in Twitter posts.

We observe that graph-based clustering performs better than LDA, in all thresholds examined. One factor that could affect the clustering results is the size of each text document. On the one hand, news articles are usually length, having a critical amount of words. On the other hand, Twitter posts are rarely larger than two lines or one sentence.

Moreover, we visualize the graph in which we present with different color each community and, at the same time, documents (posts) which are from the same topic are represented as nodes with the same shape. The visualization is shown in Figure 4.9.

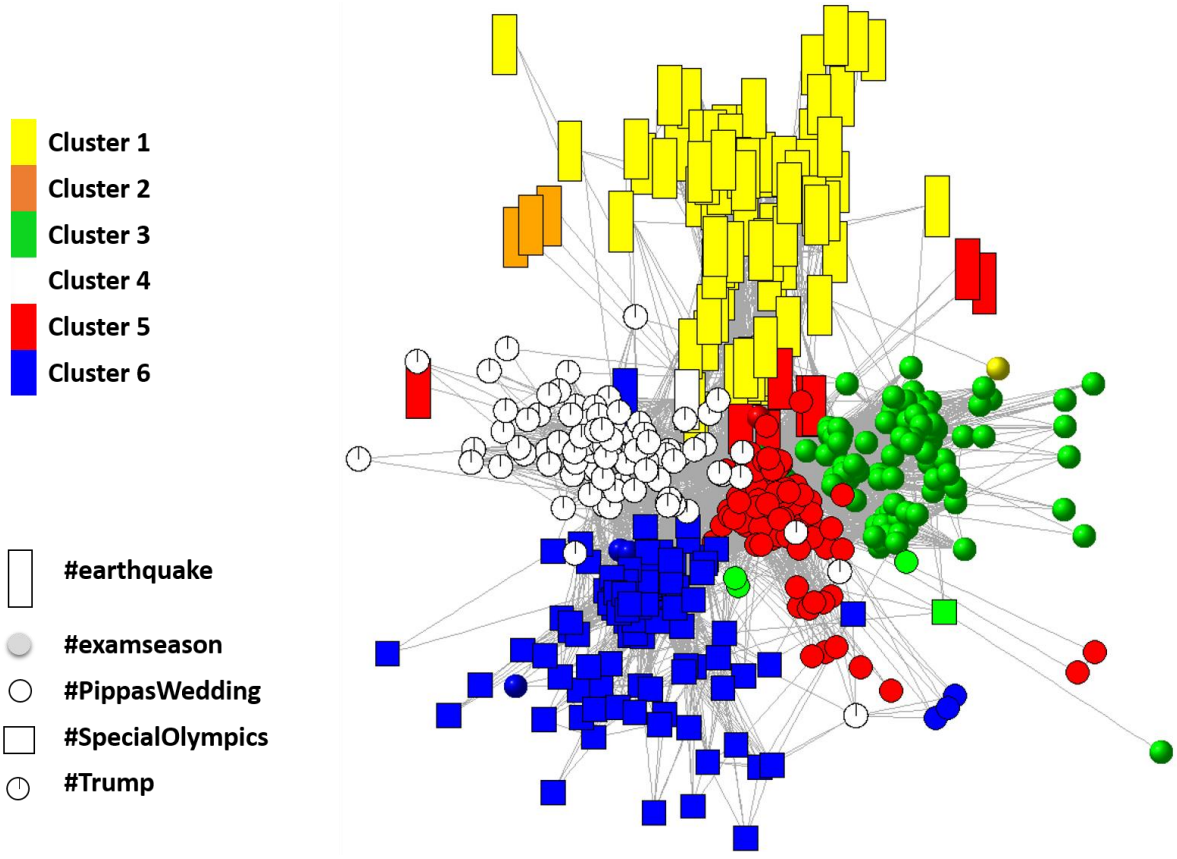


Figure 4.9: Visualization of the detected communities (color) and the ground-truth topics (shape).

In Table 4.2 we also present the confusion matrix of the variables “shape” and “color”, showing the ground-truth annotation and the detected clusters, respectively. The confusion matrix is close to being diagonal, a fact that indicates successful clustering results.

	RED	WHITE	GREEN	BLUE	YELLOW	ORANGE
CIRCLE	93	0	3	3	0	0
PIE	6	84	3	2	1	0
SPHERE	1	0	93	3	1	0
SQUARE	0	0	1	98	0	0
VRECTANGLE	8	1	0	1	79	3

Table 4.2: Confusion matrix of detected communities (color) and the ground-truth topics (shape).

Chapter 5. Bag and Graph of Visual Words

In this chapter we present the Bag of Visual Words (BoVW) model that has been used to represent an image using a statistical approach, similar to the BoW representation in text modelling. We also propose an alternative approach for the creation of visual words, but using a graph model and applying a community detection approach on the formulated graph, namely the Graph of Visual Words (GoVW). We evaluate our proposed model in two public datasets of image collections, which provide ground truth annotation for clustering purposes, and we find evidence that our method is more efficient than the BoVW model in the image clustering task.

5.1 Bag of Visual Words

In the field of Computer Vision, indexing images based on local patterns has been an active research topic over decades. The BoVW model, derived from the text indexing approach Bag-of-Words (BoW), is often used for image representation, especially when visual features are obtained without supervised methods.

The BoVW model uses local image features that are invariant to geometric and photometric transformations. The local features can be extracted from salient areas via keypoint detectors or densely sampling. Classic keypoint detectors include Harris corner detector,

maximally stable extremal region detector, affine invariant salient region detector, etc. Typical local visual descriptors correspond to keypoints detected and include the Haar descriptors, the most prominent scale-invariant feature transform (SIFT) descriptors and the histogram of oriented gradients (HOG) descriptors. Other descriptors include the gradient location and orientation histogram (GLOH) descriptor, shape descriptors, etc.

SIFT descriptors have been originally introduced in object retrieval, combined with the BoVW model (Sivic and Zisserman, 2003). Similar to the removal of stopwords in text mining, the most frequent visual words (they occur in almost all images) and very sparse terms are removed. The query and each image are represented as a sparse vector of term (visual word) occurrences, which are weighted using tf-idf scores. The similarity between the query and each image is calculated, using a distance function such as the Euclidian distance.

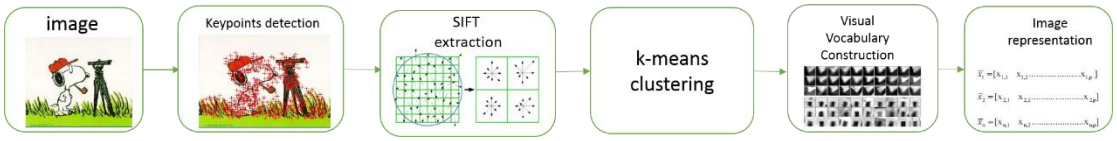


Figure 5.1: The BoVW model for image vector representation

In order to construct the visual vocabulary, we follow the framework which is presented in Figure 5.1. Firstly, keypoints are detected and, afterwards, SIFT descriptors are extracted. All SIFT descriptors from all images in the collection are aggregated and then clustered in order to provide a set of visual words (the visual vocabulary). The clustering technique is usually k-means, where the number of clusters is either tuned or set a priori to a fixed number. In addition, the inverse document frequency is used to weight the term frequencies of all visual words in all images, therefore each image is represented using tf-idf scores.

Image retrieval and image clustering are based on the BoVW model, where k-means clustering on the set of visual descriptors generate a visual vocabulary, in analogy to the vocabulary of the BoW model. In image retrieval there is a query image that is given in order to retrieve similar images in the collection. Based on the tf-idf representation of all images, it is possible to find similar-to-the-query images by minimizing the distance

between the query and each image. This distance is usually the Euclidian distance. Other applications of the BoVW model include image clustering, which aims to group together images of similar content or concept. Using the tf-idf scores of each image, the similarity of any two images is computed based on the corresponding Euclidian distances.

5.2 Graph of Visual Words

Graph-based methods have been proposed in the literature (Zhang et al., 2015) that group the set of images using a community detection method on the graph of images. The model of (Zhang et al., 2015) train images in each category and gets N descriptors per category. However, our proposed method is unsupervised and does not involve any training stage. A graph of SIFT descriptors has been proposed in (Xia and Hancock, 2009) for object indexing and retrieval purposes, but without the community detection approach that we introduce. Motivated by this graph of visual descriptors, we propose the analogue to GoW model, namely the Graph of Visual Words (GoVW), in the context of image clustering, where a graph of visual words is clustered using community detection to cluster visual descriptors into groups of visual words.

GoVW framework

The framework of Figure 5.2 presents the overall framework that we propose, in comparison to the BoVW model, which is presented in Figure 5.1. Contrary to the k-means clustering approach of all SIFT descriptors, we first estimate the first quartile of the mutual distances of all descriptors and, afterwards, using this estimation as threshold ε, we link any two SIFT descriptors whether their distance is less than ε or not:

$$l(s_\kappa, s_\lambda) = \begin{cases} 1 & if\ d(s_\kappa, s_\lambda) < \varepsilon \\ 0 & otherwise \end{cases} \tag{5.1}$$

where $s_\kappa \in \mathbb{R}^{128}$ is the κ-th SIFT descriptor, $l(s_\kappa, s_\lambda)$ is the (κ, λ) element of the adjacency matrix A of the graph of SIFT descriptors G. Community detection on the graph G provides a set C of communities $c_i, i = 1,2, \dots, n$, where n is the number of detected communities. The

community detection approach that we adopt is Louvain (Chapter 1), which is based on the maximization of modularity. Finally, using the communities as visual words, the visual vocabulary is constructed from term frequencies of appearance of visual words in images, and are weighted using tf-idf scores.

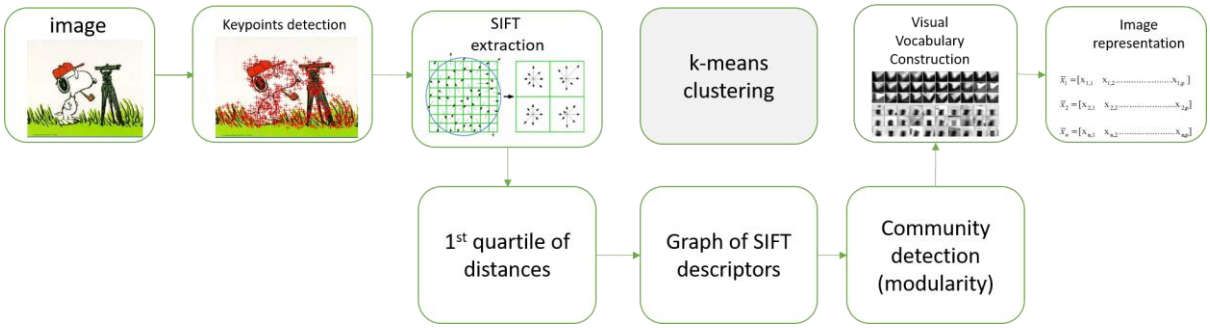


Figure 5.2: The GoVW model for image vector representation

We shall examine whether the proposed framework using the Graph of Visual Words performs better than the traditional Bag of Visual Words.

5.3 Application in Image Clustering

In this section we apply the BoVW and the GoVW models in two datasets of images, annotated with category information, i.e. it is known in which category each image belongs to. We present the datasets, the experimental settings and the final results.

Dataset description

For our experiments we use the following image collections (Gialampoukidis et al., 2016c). The Caltech 101 dataset has pictures of objects belonging to 101 categories[15], from which we get 10 images for the 8 categories (airplanes, butterfly, camera, chair, kangaroo,

[15] http://www.vision.caltech.edu/Image_Datasets/Caltech101/

saxophone, scissors, and umbrella) for the "Caltech80" dataset. The WANG dataset[16] has 1K images, belonging to 10 categories, from which we sample 10 images per category.

Settings

SIFT descriptors are extracted using the LIP-VIREO toolkit[17], where keypoints are detected using the Fast Hessian detector. For the evaluation and comparison of the two methods, we use the number of communities from the graph representation of visual words as an estimation of the number of clusters in the k-means clustering which is used in BoVW.

Results

In Table 5.1 we observe that the GoVW representation is superior to the BoVW representation model, as it is evaluated in the image clustering task.

	WANG100		**Caltech80**	
	BoV	**GoV**	**BoV**	**GoV**
NMI	0.3691	**0.4359**	0.2252	**0.3863**
VI	2.7951	**2.0081**	3.1441	**1.9936**
Adj Rand	**0.1340**	0.0864	0.0293	**0.0981**

Table 5.1: NMI, VI and Adjusted Rand evaluation for image clustering in the WANG100 and Caltech80 datasets.

NMI score increases by 6.68% in the WANG dataset and 16.11% in the Caltech dataset. Moreover, the Variation of Information index for the GoVW model is lower than the corresponding value in the BoVW model, in both datasets considered, showing that the GoVW methods obtains results that are closer to the ground truth. On the other hand, the adjusted Rand index is controversial, since it performs better in the BoVW for the WANG dataset, but for the Caltech dataset it does not show the same performance.

Apart from the quantitative evaluation of Table 5.1, we also present a qualitative evaluation of both methods in Figures 5.3 and 5.4, where we clustered a sample of the Caltech dataset,

[16] http://wang.ist.psu.edu/docs/related/
[17] LIP-VIREO toolkit: http://pami.xmu.edu.cn/~wlzhao/lip-vireo.htm

and it is shown that images showing the same object tend to belong to the same identified cluster from the GoVW model, but we do not observe the same behavior in the BoVW model.

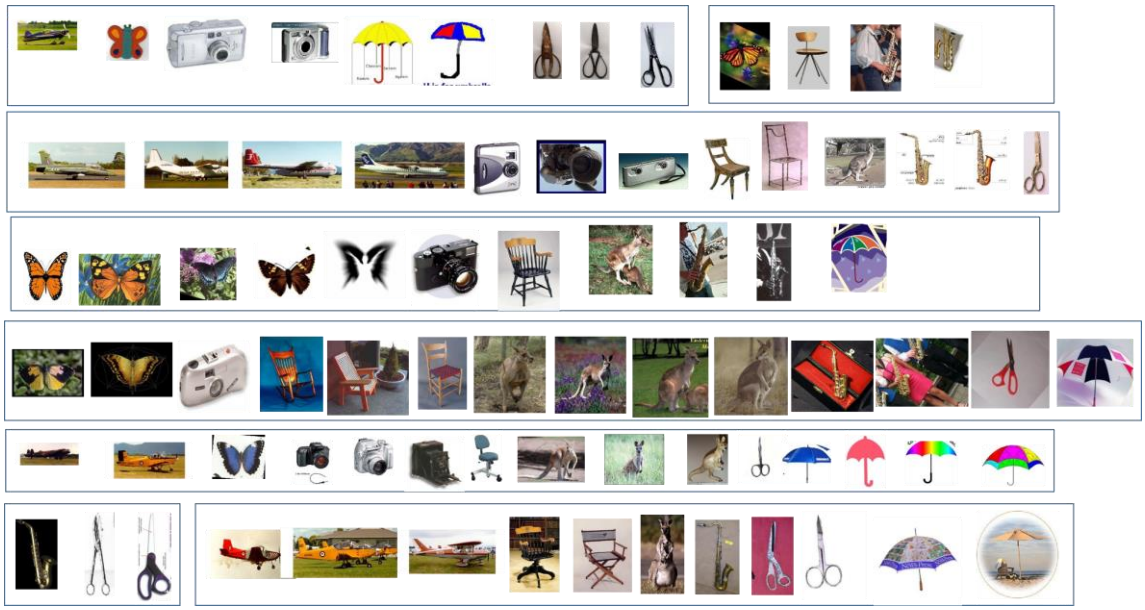


Figure 5.3: Clustering the Caltech80 dataset using the BoVW models

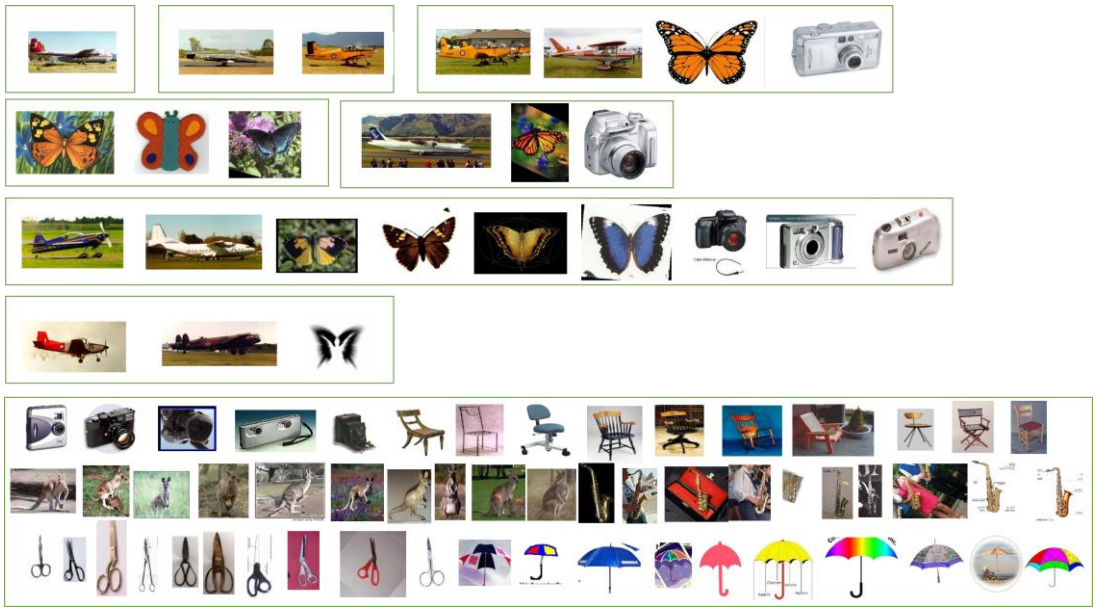


Figure 5.4: Clustering the Caltech80 dataset using the GoVW models

To sum up, the proposed GoVW model performs better than the BoVW model, replacing the k-means clustering on the set of all SIFT descriptors by a graph of SIFT descriptors formulation combined with community detection through modularity maximization. In addition, the proposed GoVW model does not require a priori knowledge of the number of visual words, hence we avoid unnecessary estimations or assumptions about the number of visual words.

In the future, we plan to further investigate the use of alternative thresholds ε in the graph formulation approach, other community detection approaches and a variety of visual descriptors. Finally, it is necessary to propose sampling techniques for the sparsification of the constructed graph of visual words, so as to have a scalable method that is applicable to larger datasets.

EPILOGUE

In this work, we have presented text and image representation using graph models, applying and evaluating our analysis in several public datasets.

We introduced a novel framework for image representation and clustering, namely the GoVW model, which is an extension of the GoW model in text representation. Contrary to the BoVW model, visual words are constructed as communities on the graph of SIFT descriptors, replacing the k-means clustering on the set of all SIFT descriptors. The proposed GoVW model performs better than the BoVW model, in the image clustering problem, as evaluated in two images collections.

One of the main advantage of the proposed GoVW model is that it does not require a priori knowledge of the number of visual words, hence it is possible to avoid unnecessary estimations or assumptions about the number of visual words. However, scalability issues need to be considered, since the computation cost becomes prohibitive for very large datasets. To that end, it is necessary to propose sampling techniques for the sparsification of the constructed graph of visual words, so as to have a scalable method that is applicable to larger datasets. As future work in the direction of the GoVW model, we plan to examine the use of alternative thresholds, community detection methods and visual descriptors.

Moreover, we have used graph-based models to extract keywords from text documents. We examined the performance of 17 keyword extraction techniques based on centrality measures and community detection approaches on the graph of words. We observed that in the case of structured text the GoW representation performs better than the simple statistical term frequency scores. On the other hand, term frequency scores were able to identify the most critical words in each document where text is less structured. We

conclude that the GoW is superior to the Bag of Words representation in the case of structured sentences that usually appear in news articles, books and technical reports.

We also proposed Mapping Entropy Closeness (MEC) centrality measure which is the second most performing keyword extraction approach, in the case of Jaccard index, following the Mapping Entropy Betweenness (MEB) scores. Centrality scores outperform community detection approaches in keyword extraction in all datasets examined. Among the centrality measures, closeness centrality performs better than the other measures. In the case of N=2, Mapping Entropy Betweenness centrality has larger Jaccard index than all other methods and among the community detection approaches, the Infomap communities contain the most important words on average and therefore obtain higher Jaccard, Average Precision and P@10.

In addition, we compared five popular text clustering methods that are based on the BoW model. In one dataset the best performing clustering method is hierarchical clustering, as indicated by all evaluation measures, while in the other two datasets LDA method is the best method in all evaluation measures apart from Rand index. This could happen because of the nature of the first dataset having only three clusters.

Regarding the evaluation by the Rand index, we observed that its highest values appear in hierarchical clustering in all cases examined. Since hierarchical clustering joins pairs of very close items it is expected that any random pair of items will have both members in the same cluster. Taking into account that Rand index is a pair counting measure (examines whether two members of random pair of items belong to the same cluster or in different ones), Rand index could be biased due to the relation between its definition and the hierarchical clustering procedure.

Although hierarchical clustering has shown better performance than LDA and other methods in the smallest dataset, it is hard to estimate the optimal height h in datasets with no ground-truth annotation. Training on an annotated sample of a dataset could serve as a solution for tuning the height parameter of a cluster dendrogram.

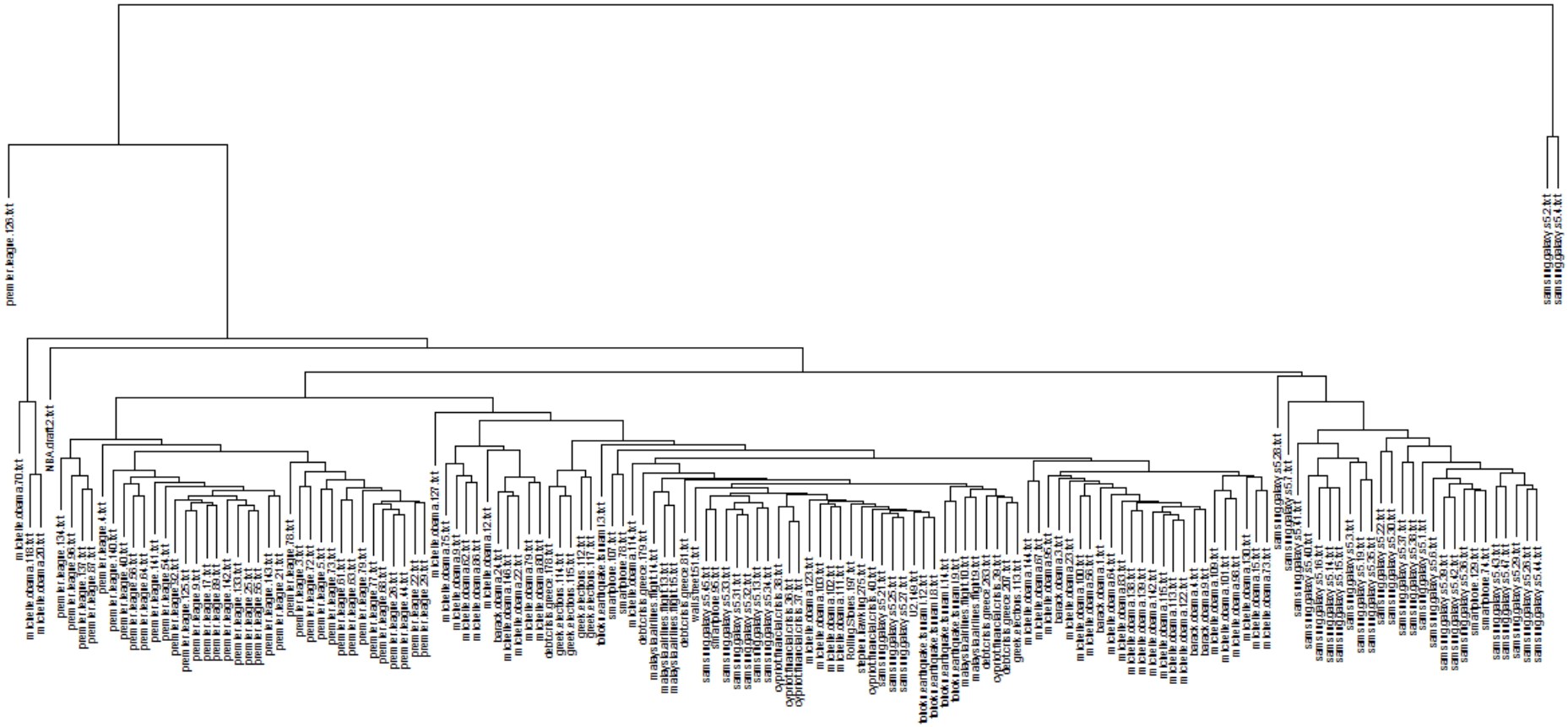
In the case of Twitter collections, the text documents involve short text of one or two sentences and graph-based clustering approaches have shown better performance than LDA.

To sum up, the proposed framework for image clustering has shown promising results, which has not appeared in the literature, to the extent of our knowledge. In the text representation of structured text documents we have shown that the use of graph-based models is superior to the statistical representation, as indicated in the keyword extraction model. Our comparative study also includes a novel centrality measure (MEC), which is in the top-3 performing methods, out of 17 approaches, and could also be investigated in other datasets with structured text.

Appendix A: tf-idf scores of the BoW model presented in Figure 2.1.

0.35	0.35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.18	0.18	0.99	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0.33	0.33	0.33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0.14	0.07	0.10	0.14	0.14	0.14	0.14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.08	0.16	0.16	0.16	0.16	0.16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.08	0	0	0	0	0	0	0	0	0	0	0	0.23	0.16	0.16	0.16	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0.06	0	0	0	0	0	0	0	0	0	0	0	0.07	0	0	0	0	0	0	0	0	0	0.14	0.14	0.14	0.14	0.14	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0.25	0.25	0	0	0	0	0
0	0	0.10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0.25	0.25	0	0
0	0	0.08	0	0	0	0	0	0.11	0.14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.20	0.20

Appendix B: The cluster dendrogram of WikiRef150 dataset



Appendix C: Sample documents from FAO and CiteULike

CiteULike [44.txt]

Exploring complex networks

Steven H. Strogatz, Department of Theoretical and Applied Mechanics and Center for Applied Mathematics, 212 Kimball Hall, Cornell University, Ithaca, New York 14853-1503, USA (e-mail: strogatz@cornell.edu)

The study of networks pervades all of science, from neurobiology to statistical physics. The most basic issues are structural: how does one characterize the wiring diagram of a food web or the Internet or the metabolic network of the bacterium Escherichia coli? Are there any unifying principles underlying their topology? From the perspective of nonlinear dynamics, we would also like to understand how an enormous network of interacting dynamical systems -- be they neurons, power stations or lasers -- will behave collectively, given their individual dynamics and coupling architecture. Researchers are only now beginning to unravel the structure and dynamics of complex networks.

Networks are on our minds nowadays. Sometimes we fear their power -- and with good reason. On 10 August 1996, a fault in two power lines in Oregon led, through a cascading series of failures, to blackouts in 11 US states and two Canadian provinces, leaving about 7 million customers without power for up to 16 hours1. The Love Bug worm, the worst computer attack to date, spread over the Internet on 4 May 2000 and inflicted billions of dollars of damage worldwide. In our lighter moments we play parlour games about connectivity. `Six degrees of Marlon Brando' broke out as a nationwide fad in Germany, as readers of Die Zeit tried to connect a falafel vendor in Berlin with his favourite actor through the shortest possible chain of acquaintances2. And during the height of the Lewinsky scandal, the New York Times printed a diagram3 of the famous people within `six degrees of Monica'. Meanwhile scientists have been thinking about networks too. Empirical studies have shed light on the topology of food webs4,5, electrical power grids, cellular and metabolic networks6-9, the World-Wide Web10, the Internet backbone11, the neural network of the nematode worm Caenorhabditis elegans12, telephone call graphs13, coauthorship and citation networks of scientists14-16, and the quintessential `old-boy' network, the overlapping boards of directors of the largest companies in the United States17 (Fig. 1). These databases are now easily accessible, courtesy of the Internet. Moreover, the availability of powerful computers has made it feasible to probe their structure; until recently, computations involving million-node

networks would have been impossible without specialized facilities. Why is network anatomy so important to characterize? Because structure always affects function. For instance, the topology of social networks affects the spread of information and disease, and the topology of the power grid affects the robustness and stability of power transmission. From this perspective, the current interest in networks is part of a broader movement towards research on complex systems. In the words of E. O. Wilson18, "The greatest challenge today, not just in cell biology and ecology but in all of science, is the accurate and complete description of complex systems. Scientists have broken down many kinds of systems. They think they know most of the elements and forces. The next task is to reassemble them, at least in mathematical models that capture the key properties of the entire ensembles."

But networks are inherently difficult to understand, as the following list of possible complications illustrates. 1. Structural complexity: the wiring diagram could be an intricate tangle (Fig. 1). 2. Network evolution: the wiring diagram could change over time. On the World-Wide Web, pages and links are created and lost every minute. 3. Connection diversity: the links between nodes could have different weights, directions and signs. Synapses in

Dynamical systems can often be modelled by differential equations dx/dt v(x), where x(t) (x1(t), ..., xn(t)) is a vector of state variables, t is time, and v(x) (v1(x), ..., vn(x)) is a vector of functions that encode the dynamics. For example, in a chemical reaction, the state variables represent concentrations. The differential equations represent the kinetic rate laws, which usually involve nonlinear functions of the concentrations. Such nonlinear equations are typically impossible to solve analytically, but one can gain qualitative insight by imagining an abstract n-dimensional state space with axes x1, ..., xn. As the system evolves, x(t) flows through state space, guided by the `velocity' field dx/dt v(x) like a speck carried along in a steady, viscous fluid. Suppose x(t) eventually comes to rest at some point x. Then the velocity must be zero there, so we call x* a fixed point. It corresponds to an equilibrium state of the physical system being modelled. If all small disturbances away from x* damp out, x* is called a stable fixed point -- it acts as an attractor for states in its vicinity. Another long-term possibility is that x(t) flows towards a closed loop and eventually circulates around it forever. Such a loop is called a limit cycle. It represents a self-sustained oscillation of the physical system. A third possibility is that x(t) might settle onto a strange attractor, a set of states on which it wanders forever, never stopping or repeating. Such erratic, aperiodic motion is considered chaotic if two nearby states flow away from each other exponentially fast. Long-term prediction is impossible in a real chaotic system because of this exponential amplification of small uncertainties or measurement errors*

NATURE | VOL 410 | 8 MARCH 2001 | www.nature.com

FAO [00950e.txt]

Where to purchase FAO publications locally - Points de vente des publications de la FAO - Puntos de venta de publicaciones de la FAO
· ANGOLA
Empresa Nacional do Disco e de Publicaηues, ENDIPU-U.E.E.
Rua Cirilo da Conceiηyo Silva, N° 7
C.P. N° 1314-C, Luanda
· ARGENTINA
Librerva Agropecuaria
Pasteur 743, 1028 Buenos Aires
Oficina del Libro Internacional
Av. Cordoba 1877, 1120 Buenos Aires
E-mail: olilibro@satlink.com
· AUSTRALIA
Hunter Publications
P.O. Box 404, Abbotsford, Vic. 3067
Tel.:(03) 9417 5361
Fax: (03) 914 7154
E-mail: jpdavies@ozemail.com.au
· AUSTRIA
Gerold Buch & Co.
Weihburggasse 26, 1010 Vienna
· BANGLADESH
Association of Development Agencies in Bangladesh
House No. 1/3, Block F,
Lalmatia, Dhaka 1207
· BELGIQUE
M.J. De Lannoy
202, avenue du Roi, 1060 Bruxelles
CCP 000-0808993-13
E-mail: jean.de.lannoy@infoboard.be
· BOLIVIA
Los Amigos del Libro
Av. Herovnas 311, Casilla 450
Cochabamba;

Mercado 1315, La Paz
· BOTSWANA
Botsalo Books (Pty) Ltd
P.O. Box 1532, Gaborone
· BRAZIL
Fundaηγo Getϋlio Vargas
Praia do Botafogo 190, C.P. 9052
Rio de Janeiro
E-mail: valeria@sede.fgvrj.br
Nϋcleo Editora da Universidade Federal Fluminense
Rua Miguel de Frias 9
Icarav-Niterσi 24
220-000 Rio de Janeiro
Fundaηγo da Universidade Federal do Paranα - FUNPAR
Rua Alfredo Bufrem 140, 30° andar
80020-240 Curitiba
· CAMEROON
CADDES
Centre Africain de Diffusion et
Dιveloppement Social
B.P. 7317 Douala Bassa
Tel.: (237) 43 37 83
Fax: (237) 42 77 03
· CANADA
Renouf Publishing
5369 chemin Canotek Road, Unit 1
Ottawa, Ontario K1J 9J3
Tel.: (613) 745-2665
Fax: (613) 745 7660
Website: www.renoufbooks.com
E-mail: renouf@fox.nstn.ca
· CHILE
Librerva - Oficina Regional FAO
c/o FAO Officina Regional para Amιrica
Latina y el Caribe (RLC)

Bibliography

Abilhoa, W. D., & De Castro, L. N. (2014). A keyword extraction method from twitter messages represented as graphs. *Applied Mathematics and Computation*, *240*, 308-325.

Beliga, S., Meštrović, A., & Martinčcić-Ipšić, S. (2014). Toward selectivity based keyword extraction for Croatian news. *arXiv preprint arXiv:1407.4723*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Boudin, F. (2013, October). A comparison of centrality measures for graph-based keyphrase extraction. In *International Joint Conference on Natural Language Processing (IJCNLP)* (pp. 834-838).

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, *25*(2), 163-177.

Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, *56*(18), 3825-3833.

Campello, R. J., Moulavi, D., & Sander, J. (2013, April). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 160-172). Springer Berlin Heidelberg.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., & Charrad, M. M. (2014). Package 'NbClust'. *J. Stat. Soft*, *61*, 1-36.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, *70*(6), 066111.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Danon, Leon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. "Comparing community structure identification." *Journal of Statistical Mechanics: Theory and Experiment* 2005, no. 09 (2005): P09008.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. *Biometrics*, *21*(3), 768-769.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, *486*(3), 75-174.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35-41.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social networks*, *1*(3), 215-239.

Gantmakher, F. R. (1998). *The theory of matrices* (Vol. 131). American Mathematical Soc..

Gialampoukidis, I., Kalpakis, G., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2016a). Key player identification in terrorism-related social media networks using centrality measures. In *European Intelligence and Security Informatics Conference (EISIC 2016), August* (pp. 17-19).

Gialampoukidis, I., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2016b). Community detection in complex networks based on DBSCAN* and a Martingale process. In *Semantic and Social Media Adaptation and Personalization (SMAP), 2016 11th International Workshop on* (pp. 1-6). IEEE.

Gialampoukidis, I., Vrochidis, S., & Kompatsiaris, I. (2016c). Fast visual vocabulary construction for image retrieval using skewed-split kd trees. In *International Conference on Multimedia Modeling* (pp. 466-477). Springer International Publishing.

Gialampoukidis, I., Vrochidis, S., & Kompatsiaris, I. (2016d). A Hybrid framework for news clustering based on the DBSCAN-Martingale and LDA. In *Machine Learning and Data Mining in Pattern Recognition* (pp. 170-184). Springer International Publishing.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), 7821-7826.

Grineva, M., Grinev, M., & Lizorkin, D. (2009, April). Extracting key terms from noisy and multitheme documents. In *Proceedings of the 18th international conference on World wide web* (pp. 661-670). ACM.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100-108.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification*, *2*(1), 193-218.

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, *40*(9), 1098-1101.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241-254.

Kulis, B., & Jordan, M. I. (2011). Revisiting k-means: New algorithms via Bayesian nonparametrics. *arXiv preprint arXiv:1111.0352*.

Lahiri, S., Choudhury, S. R., & Caragea, C. (2014). Keyword and keyphrase extraction using centrality measures on collocation networks. *arXiv preprint arXiv:1401.6571*.

Lloyd, S. (1957, 1982). Least squares quantization in PCM. *IEEE transactions on information theory*, *28*(2), 129-137.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

Meilă, M. (2003). Comparing clusterings by the variation of information. In *Learning theory and kernel machines* (pp. 173-187). Springer Berlin Heidelberg.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), 026113.

Nie, T., Guo, Z., Zhao, K., & Lu, Z. M. (2016). Using mapping entropy to identify node centrality in complex networks. *Physica A: Statistical Mechanics and its Applications*, *453*, 290-297.

Pons, P., & Latapy, M. (2005, October). Computing communities in large networks using random walks. In *International Symposium on Computer and Information Sciences* (pp. 284-293). Springer Berlin Heidelberg.

Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, *76*(3), 036106.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, *66*(336), 846-850.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, *3*(4), 333-389.

Rosvall, M., & Bergstrom, C. T. (2007). *Maps of information flow reveal community structure in complex networks*. Technical report.

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118-1123.

Rosvall, M., & Bergstrom, C. T. (2010). Mapping change in large networks. *PloS one*, *5*(1), e8694.

Rousseau, F., & Vazirgiannis, M. (2013, October). Graph-of-word and TW-IDF: new approach to ad hoc IR. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 59-68). ACM.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*(4), 581-603.

Sivic, J., & Zisserman, A. (2003, October). Video google: A text retrieval approach to object matching in videos. In *iccv* (Vol. 2, No. 1470, pp. 1470-1477).

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2004, December). Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. In *NIPS* (pp. 1385-1392).

Tsatsaronis, G., Varlamis, I., & Nørvåg, K. (2010, August). SemanticRank: ranking keywords and sentences using semantic graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 1074-1082). Association for Computational Linguistics.

Xia, S., & Hancock, E. R. (2009, May). Pairwise similarity propagation based graph clustering for scalable object indexing and retrieval. In International Workshop on Graph-Based Representations in Pattern Recognition (pp. 184-194). Springer Berlin Heidelberg.

Xie, Z. (2005, June). Centrality measures in text mining: prediction of noun phrases that appear in abstracts. In *Proceedings of the ACL student research workshop* (pp. 103-108). Association for Computational Linguistics.

Zhang, W., Lu, H., Sun, S., & Gu, X. (2015, November). A Graph Community and Bag of Categorized Visual Words Based Image Retrieval. In *International Conference on Neural Information Processing* (pp. 473-480). Springer International Publishing.