Belief Detection and Temporal Analysis of Experts in Question Answering Communities: case study Stack Overflow

Dorra Attiaoui

ANNÉE 2017

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

En Cotutelle Internationale avec
L'université de Tunis, Tunisie

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

Ecole doctorale MATHSTIC

présentée par

Dorra Attiaoui

préparée à l'unité de recherche UMR 6074 IRISA
Institut de Recherche en Informatique et Système Aléatoire
Université de Rennes I

Belief Detection and Temporal Analysis of Experts in Question Answering Communities: case study Stack Overflow

Thèse soutenue à l'Université de Rennes 1 le 01/12/2017

devant le jury composé de :

Francis ROUSSEAUX
Professeur à l'Université de Reims, France / Examinateur

Julien VELCIN
Maitre de conférences, HDR, Université de Lyon 2, France / Rapporteur

Allal HADJALI
Professeur à l'Université de Poitiers, France / Rapporteur

Boutheina BEN YAGHLANE
Professeur à l'Universités de Carthage, Tunisie / co-directrice de thèse

Arnaud MARTIN
Professeur à l'Université de Rennes 1, France / directeur de thèse

Abstract

During the last decade, people have changed the way they seek information online. Between question answering communities, specialized websites, social networks, the Web has become one of the most widespread platforms for information exchange and retrieval. Question answering communities provide an easy and quick way to search for information needed in any topic. The user has to only ask a question and wait for the other members of the community to respond. Any person posting a question intends to have accurate and helpful answers. Within these platforms, we want to find experts. They are key users that share their knowledge with the other members of the community. Expert detection in question answering communities has become important for several reasons such as providing high quality content, getting valuable answers, etc.

In this thesis, we are interested in proposing a general measure of expertise based on the theory of belief functions. Also called the mathematical theory of evidence, it is one of the most well known approaches for reasoning under uncertainty.

In order to identify experts among other users in the community, we have focused on finding the most important features that describe every individual. Next, we have developed a model founded on the theory of belief functions to estimate the general expertise of the contributors. This measure will allow us to classify users and detect the most knowledgeable persons.

Therefore, once this metric defined, we look at the temporal evolution of users' behavior over time. We propose an analysis of users activity for several months in community. For this temporal investigation, we will describe how do users evolve during their time spent within the platform. Besides, we are also interested on detecting potential experts during the beginning of their activity. The effectiveness of these approaches is evaluated on real data provided from Stack Overflow.

Keywords: Question Answering Communities, Expertise, Experts detection, Clustering, Theory of Belief Functions, Combination

Résumé

L'émergence du Web 2.0 a changé la façon avec laquelle les gens recherchent et obtiennent des informations sur internet. Entre sites communautaires spécialisés, réseaux sociaux, l'utilisateur doit faire face à une grande quantité d'informations. Les sites communautaires de questions réponses représentent un moyen facile et rapide pour obtenir des réponses à n'importe quelle question qu'une personne se pose. Tout ce qu'il suffit de faire c'est de déposer une question sur un de ces sites et d'attendre qu'un autre utilisateur lui réponde. Dans ces sites communautaires, nous voulons identifier les personnes très compétentes. Ce sont des utilisateurs importants qui partagent leurs connaissances avec les autres membres de leurs communauté. Ainsi la détection des experts est devenue une tache très importantes, car elle permet de garantir la qualité des réponses postées sur les différents sites.

Dans cette thèse, nous proposons une mesure générale d'expertise fondée sur la théorie des fonctions de croyances. Cette théorie nous permet de gérer l'incertitude présente dans toutes les données émanant du monde réel.

D'abord et afin d'identifier ces experts parmi la foule d'utilisateurs présents dans la communauté, nous nous sommes intéressés à identifier des attributs qui permettent de décrire le comportement de chaque individus. Nous avons ensuite développé un modèle statistique fondé sur la théorie des fonctions de croyance pour estimer l'expertise générale des usagers de la plateforme. Cette mesure nous a permis de classifier les différents utilisateurs et de détecter les plus experts d'entre eux.

Par la suite, nous proposons une analyse temporelle pour étudier l'évolution temporelle des utilisateurs pendant plusieurs mois. Pour cette partie, nous décrirons comment les différents usagers peuvent évoluer au cours de leur activité dans la plateforme. En outre, nous nous sommes également intéressés à la détection des experts potentiels pendant les premiers mois de leurs inscriptions dans un site. L'éfficacité de ces approches a été validée par des données réelles provenant de Stack Overflow.

Mots clés: Réseaux Communautaires de Questions Réponses, Expertise, Détection des Experts, Clustering, Théorie des fonctions de croyance, Combinaison

Acknowledgements

"*En vérité, le chemin importe peu, la volonté d'arriver suffit à tout.*"
Albert Camus, Le Mythe de Sisyphe

So many people have supported me and encouraged me to finish my PhD.
I would like to express my deepest thanks to my advisors Pr. Arnaud Martin (University of Rennes 1, France) and Pr. Boutheina Ben Yaghlane (University of Carthage, Tunisia) for their guidance and encouragement during these five years, from the Master degree to the defense of my thesis.

A big thank you to Arnaud Martin for his patience, his competence, his continuous support and his modesty. It was a real pleasure to work with him, I have learned so much.

Second, I would like to thank the members of my thesis committee: Pr. Francis Rousseaux (University of Reims, France), Pr. Allal Hadjali (University of Poitiers, France), and Dr. Julien Velcin (University of Lyon 2, France) for accepting to review my dissertation.

A very special thank to my colleagues in DRUID Team and the IT department of IUT Lannion for help, kindness and support: especially Mouloud, Claude, Tassadit, Denis, Anne-Isabelle, J-C ... It was an honor to have the opportunity to work and learn from each one of you

Last but not least, my parents, without whom nothing would have been possible. I would like to tell them "*Soyons reconnaissants aux personnes qui nous donnent du bonheur, elles sont les charmants jardiniers par qui nos âmes sont fleuries*". My beloved sister Nadia & co and little sestra Yosra for their patience (sorry for complaining so much girls !). My three big little nieces, for their foolishness and their laughs. My big brother Walid & co for their support.

Moreover, one important gift of life is friendship, and I have received it, so to all my friends thank you (especially my best friend, "my friend" and "my friends of misery"). (To avoid the risk of forgetting some names, I will not cite any of them).

The thesis has been an adventure on a human and intellectual level.

Thank you to everyone who **believed** in me...

Contents

List of Tables

List of Figures

Abbreviations and notations

In the following, a list as exhaustive as possible of abbreviations and notations used in this thesis:

Theory of Belief Functions

- Ω: is the frame of discernment;
- $\omega_1, \omega_2, \ldots, \omega_n$: hypothesis in Ω; they are singletons;
- m: is a mass function defined on any frame of discernment Ω;
- ${}^{\alpha}m$: discounted mass function;
- m_{X*}: categorical mass function;
- m_{Ω}: vacuous mass function;
- p: probability;
- $BetP$: pignistic probability;
- $\oplus$: operator of the Dempster's combination rule;
- m_D: mass resulting from the Dempster's combination rule;
- m_{conj}: mass resulting from the conjunctive combination rule;
- m_{disj}: mass resulting from the disjunctive combination rule;

Data Analysis

- HCA: Hierarchical Cluster Analysis;
- PCA: Principle Component Analysis;
- k correlation index;
- p number of original variables;
- AEC: Average Eigenvalue Criterion;

- $CAEC$: Corrected Average Eigenvalue Criterion;
- $DownV$: DownVote;
- $NbQu$: number of questions;
- $NbAn$: number of answers;
- $NbAccAn$: number of accepted answers;

Expertise detection

- AV_i: number of votes gained by answers posted;
- QV_i: number of votes gained by questions asked;
- E: Expert;
- A: Apprentice;
- O: Occasionnal;
- BME: Belief Measure of Expertise;
- α^T: discounting time;
- nd_i: number of days;
- CI: confidence interval;
- SE: standard error;
- $TBME$: Temporal Belief Measure of Expertise;
- CDF: Cumulative Distribution Functions;

Clustering

- C_i: class of index i;
- N_c: number of classes;
- n_i: number of elements in a class C_i;
- $Intra$: Intra class inertia;
- $Inter$: Inter class inertia;
- c_G: gravity center of a class;

- S: Silhouette;
- $a(i)$: mean distance between an object and its peers in the same class;
- $b(i)$: mean distance between an object and the other objects of the closest class.l;
- $I(C_i)$: mean of the distances between the objects of a class and its center;
- SSB: overall between-cluster variance;
- SSW: overall within- cluster variances;
- $d(C_i, C_j)$: distance between cluster C_i and C_j;
- y: observed value;
- $\overline{y}$: mean value of the observations;
- $\hat{y}$: fitted value of the observation;

1 Détection et Analyse temporelle des experts dans les réseaux communautaires de questions réponses : étude de cas Stack Overflow

1 Introduction

L'émergence du Web 2.0 durant la dernière décennie s'est faite grâce aux réseaux sociaux avec lesquels nous pouvons interagir et communiquer avec n'importe quelle personne dans le monde entier. Avec plus de 3.835 milliards d'individus connectés à Internet en 2017, cela a changé notre façon de communiquer et chercher des informations.

Parmi les nouvelles tendances du moment le recours aux forums spécialisés pour obtenir des informations précises concernant un thème bien déterminé s'est démocratisé. On retrouve différentes plateformes dans différents domaines, où les membres viennent poser des questions et y répondre. Ce genre d'outils peut aussi bien être à usage professionnel (Stack Overflow[1], Quora[2]) ou non professionnels (Yahoo!Answers[3], TripAdvisor[4], etc.)

Il existe deux types d'utilisateurs : les gens qui viennent en tant que demandeurs pour trouver des réponses à leurs questions, et des personnes compétentes qui répondent aux questions des usagers les moins experts. La force de ces réseaux réside dans l'expertise de certains de ses usagers.

Dans ce travail, nous nous intéressons en particulier à l'identification de ces experts. Ils sont certes peu nombreux dans ce type de plateformes mais ce sont les utilisateurs clés au sein de leurs communautés. Ce sont eux qui garantissent la qualité des échanges dans les communautés, et donc la qualité de la plateforme.

[1] https://stackoverflow.com/
[2] https://www.quora.com/
[3] https://fr.answers.yahoo.com/
[4] https://www.tripadvisor.fr/

La détection des experts a fait l'objet de plusieurs recherches scientifiques. Ainsi, certains chercheurs ont examiné le comportement des différents utilisateurs. Ils ont étudié la motivation de ces individus ainsi que leur capacité à aider les autres.

Stack Overflow a mis en place en Septembre 2012 un système de réputation pour les participants de cette plateforme. Pour certains chercheurs, cette mesure est une indication de l'expertise d'une personne (Movshovitz-Attias et al., 2013).

Cependant, dans cette thèse nous considérons cette métrique comme étant plus un indicateur de popularité que d'expertise. Comme nous le montrons, elle reflète comment la communauté perçoit un utilisateur.

(Kasneci et al., 2011) ont identifié trois niveaux incertitude dans ces sites, le premier est lié à l'extraction et l'intégration des données, le second aux sources d'informations et finalement aux informations elles-mêmes. Dans cette thèse, nous nous intéressons uniquement au deuxième point correspondant aux utilisateurs et plus précisément aux experts.

Afin de palier aux problèmes liés à l'imperfection des données, de nombreuses théories ont vu le jour, telles que la théorie des possibilités, des probabilités, ou encore la théorie des fonctions de croyances. Dans cette thèse notre choix s'est porté sur la théorie des fonctions de croyance. Cette dernière nous offre un cadre mathématique très riche qui permet de représenter différents types d'imperfections. Elle permet aussi la combinaison des données et la gestion du conflit qui peut en résulter. Récemment, cette théorie a été adoptée dans plusieurs travaux reliés aux réseaux sociaux tels que (Dlala et al., 2015) ou encore (Jendoubi et al., 2017).

2 Détection des experts dans les communautés

Afin d'identifier des experts dans les communautés en ligne, nous pouvons distinguer deux approches majeures : les approches fondées sur le classement et les approches fondées sur les attributs comme présenté par (Sahu et al., 2016).

Alors que les approches fondées sur le classement visent à calculer une métrique ou un score par utilisateur qui sert à séléctionner un nombre spécifique d'utilisateurs qualifiés d'experts. Les approches fondées sur les attributs visent à identifier un certain nombre de caractéristiques qui décrivent chaque utilisateur. Ces attributs permettent de classer les utilisateurs comme experts ou non spécialisés en fonction des techniques d'apprentissage. Les deux approches s'appuient sur des connaissances antérieures à partir de l'ensemble de données manipulées.

Le problème principal avec toutes les approches fondées sur le classement est que le nombre d'experts est défini au début de chaque méthode. Dans une communauté regroupant des milliers voire des centaines de milliers d'individus, nous ne pouvons pas préciser dès le début le nombre de personnes que nous considérons comme expertes. En outre, certaines des méthodes peuvent faire l'impasse sur certaines caractéristiques

importantes des utilisateurs comme le nombre de meilleures réponses fournies, le temps d'activité, etc.

De ce fait, la deuxième approche fondée sur les attributs propose une bonne alternative qui permet de mesurer l'expertise des utilisateurs d'une façon plus globale. Elles prennent en considération différents éléments qui décrivent le comportement des usagers des plateformes.

3 Théorie des fonctions de croyance

Les théories de l'incertain sont issues de la précision des mathématiques classiques et de l'imprécision émanant du monde réel. Plusieurs études ont permis d'aboutir à la théorie des ensembles flous ou des fonctions de croyance permettant de représenter l'imprécision et l'incertitude des connaissances.

La théorie des fonctions de croyance initialement introduite par (Dempster, 1967), formalisée ensuite dans les travaux de (Shafer, 1976) a été utilisée dans différentes applications telles que la classification ainsi que le traitement d'images (Khaleghi et al., 2013), réseaux sociaux (Attiaoui et al., 2015), (Zhou et al., 2016)...
La fusion d'informations permet d'aider les décideurs qu'ils soient humains ou logiciels à la prise de décision en réduisant les données. Cette fusion fournit un résultat facilement interprétable surtout quand il s'agit de domaines où le nombre de sources est considérable comme c'est le cas pour réseaux sociaux...

3.1 Formalisme

A partir d'un cadre de discernement Ω ($\Omega = \{\omega_1, ..., \omega_n\}$) qui est l'ensemble de toutes les hypothèses, nous définissons une fonction de masse sur l'ensemble de tous les sous ensembles possibles de Ω à qui on affecte une valeur comprise entre $[0, 1]$ représentant ainsi sa masse de croyance élémentaire. Formellement une fonction de masse m est définie comme suit :

$$m : 2^{\Omega} \mapsto [0, 1]. \tag{1.1}$$

$$\sum_{X \subseteq \Omega} m(X) = 1. \tag{1.2}$$

3.2 Les règles de combinaison

Dans la littérature, différentes règles de combinaison ont été proposées pour effectuer la fusion d'information.

3.2.1 La règle de Dempster

La première règle de combinaison a été introduite par Dempster en 1967. Etant donné deux fonctions de masses m_1 and m_2, pour tout $X \in 2^\Omega$, $X \neq \emptyset$, la règle de Dempster est définie par :

$$m_D(X) = \frac{1}{1-k} \sum_{Y_1 \cup Y_2 = X} m_1(Y_1) m_2(Y_2) \tag{1.3}$$

où $k = \sum_{Y_1 \cap Y_2 = \emptyset} m_1(Y_1) m_2(Y_2)$ est l'inconsistance émanant de la combinaison aussi appelé *conflit global*. La valeur $1 - k$ est un facteur de normalisation de la règle de combinaison.

3.2.2 La règle comjonctive

La règle de combinaison conjonctive proposée par (Smets, 1990) est utilisée lorsque les sources d'information sont considérées comme étant fiables et indépendantes cognitivement. Elle est donnée pour tout $X \in 2^\Omega$ par :

$$m_{conj}(X) = \sum_{Y_1 \cap Y_2 = X} m_1(Y_1) m_2(Y_2) \tag{1.4}$$

3.3 Prise de décision

Pour la prise de décision, la transformation pignistique, proposée par (Smets, 2005), permet de transformer les fonctions de masse en mesures de probabilité. Elles permettent ainsi une prise de décision sur seulement des singletons. La probabilité pignistique définie par :

$$BetP(X) = \sum_{Y \in 2^\Omega, Y \neq \emptyset} \frac{|X \cap Y|}{|Y|} * \frac{m(Y)}{1 - m(\emptyset)}, \forall X \in 2^\Omega, X \neq \emptyset \tag{1.5}$$

4 Analyse des données

La base de données manipulée a été téléchargée via le site de stockage d'archives de Stack Overflow[5].

Une étape très importante est la caractérisation des utilisateurs qui consiste à extraire des attributs pour représenter chacun des usagers de la plateforme. Lors de cette

[5] https://archive.org/download/stackexchange

étape, nous avons procédé à une analyse des données. Nous nous sommes retrouvés face à un grand nombre de variables à prendre en compte pour décrire le comportement des utilisateurs. Afin de mieux comprendre cela nous avons eu recours à une analyse en composantes principales. Par la suite nous avons effectué une classification mixte qui est utilisée quand le nombre d'individus est très grand. Cette méthode de classification se compose de deux étapes : une classification non supervisée et une classification hiérarchique ascendante. Grâce à ces différentes méthodes d'analyse, nous avons pu caractériser trois types d'utilisateurs présents dans Stack Overflow :

- **Les occasionnels** : ce sont des utilisateurs peu actifs dans la communauté. Ils postent des questions de temps à autre seulement quand ils sont à la recherche d'information.
- **Les apprentis** : ce sont des utilisateurs actifs. Ils souhaitent avoir de la reconnaissance au sein de leur communauté. Ils postent essentiellement beaucoup de réponses mais malheureusement la qualité n'est toujours pas garantie.
- **Les experts** : ces utilisateurs ont de grandes connaissances dans un ou plusieurs domaines. Ils fournissent des réponses qui sont sélectionnées comme étant les meilleures. Ils garantissent la qualité des échanges au sein de leurs communautés.

Prendre en considération l'incertitude lors de la caractérisation des utilisateurs permet d'améliorer l'identification des experts dans les réseaux communautaires.

5 Détection des experts avec la théorie des fonctions de croyance

Dans cette partie nous allons détailler la démarche à suivre pour la classification des utilisateurs et la détection des experts dans Stack Overflow.

5.1 Définition des attributs

- **Nombre de votes positifs** : la somme des votes positifs récoltés en postant des questions et des réponses.
- **Nombre de votes négatifs**: la somme des votes négatifs récoltés en posant des questions et des réponses.
- **Temps d'activité :** le nombre de jours d'activités de l'utilisateur depuis son inscription.
- **Nombre de questions posées** : nombre de questions posées dans la plateforme.
- **Nombre de réponses postées** : nombre de réponses postées dans la plateforme.

- **Nombre de meilleures réponses** : nombre de réponses choisies comme étant les meilleures.

5.2 Modélisation

- Si un utilisateur a un grand score positif relié aux réponses données, cette personne peut être experte. Dans le cas contraire, cette personne sera sanctionnée par des votes négatifs qui feront baisser ce score.

- Si un utilisateur a un score élevé relatif aux questions, cela pourrait signifier que cette personne est un apprenti à la recherche d'informations. Elle sera récompensée par la communauté pour avoir posté des questions bien formulées et fort intéressantes. Dans le cas contraire, ce score sera faible.

- Si un utilisateur a un nombre élevé de réponses postées, cela peut être justifié par deux faits. Tout d'abord, cette personne est considérée comme étant experte, fournissant un contenu de haute qualité. Deuxièmement, cette personne peut être un apprenti essayant de devenir un expert en prouvant à la communauté qu'il/elle peut être aussi fiable qu'un expert.

- Si un utilisateur a un nombre élevé de questions postées, cela peut représenter un expert ou un apprenti. Les deux posent beaucoup de questions d'après l'analyse faite précédemment.

- Si un utilisateur a un grand nombre de réponses acceptées comme étant les meilleures, ceci ne peut signifier qu'une chose, c'est que cette personne est experte.

5.3 Classification des utilisateurs dans Stack Overflow

Afin de classer les utilisateurs dans un des trois groupes définis précédemment, nous proposons de calculer leur degré d'expertise. Cette mesure est le résultat de la combinaison de Dempster sur des fonctions de masses prédéfinies selon le modèle décrit dans la section 5.2. Cette mesure est aussi affaiblie avec le temps d'activité de chaque utilisateur.
Une fois les informations caractérisant les usagers fusionnées, nous utilisons la probabilité pignistique de l'équation (5) pour procéder à la prise de décision. En d'autres termes le choix de la classe à laquelle l'utilisateur appartient.

La valeur de la mesure d'expertise définie dans cette thèse se situe dans un intervalle entre $[0, 1]$. Plus la personne est experte, plus cette valeur est proche de 1. Si l'utilisateur est classé comme occasionnel, la valeur de sa mesure tend ou vers 0. Nous avons comparé nos résultats avec ceux de la réputation de Stack Overflow, et une Mixture de

Gaussiennes. Nous avons trouvé que notre mesure était plus performante que les deux autres approches pour la détection des experts dans ce type de communauté.

6 Analyse temporelle des utilisateurs et prédiction des experts potentiels

Dans le but d'étudier le comportement des utilisateurs au cours du temps, nous procédons à une analyse temporelle. Pour cela nous définissions pour chaque mois les différents attributs de chaque usager et ce pendant une période de 15 mois.

Cependant, avant de considérer l'aspect temporel, nous nous sommes focalisés sur l'identification des experts potentiels. Ce sont des personnes très importantes dans la communauté car ce sont les futurs experts qui animeront la communauté et veilleront à la qualité des échanges sur la plateforme. Il faut les détecter dès les premiers mois après leur inscription. Il sont peu nombreux, ce qui fait de cette étape une tache difficile. Pour cela nous avons comparé leurs degrés d'expertises généraux par rapport à leur expertise au bout de 100 jours d'activité dans la plateforme. Par la suite, nous avons mesuré l'évolution de l'expertise des utilisateurs pendant 15 mois consécutifs. Ceci nous a permis d'avoir une vision générale de la motivation et de l'activité des usagers pendant plus d'une année. Certains utilisateurs gardent le même comportement pendant des mois et ne changent pas. Mais d'autres évoluent. Nous avons ainsi pu voir la progression de certains utilisateurs qui ont acquis des connaissances et sont devenus experts grâce à leur participation au sein de la communauté.

Conclusion

Dans cette thèse, nous nous sommes concentrés sur la détection des experts dans les réseaux communautaires de questions/réponses. Nous avons proposé une analyse des données de chaque utilisateur. Par la suite nous avons défini une mesure d'expertise fondée sur la théorie des fonctions de croyances. Cette métrique a permis de caractériser les usagers de Stack Overflow en trois catégories : Experts, Apprentis et Occasionnels. Nous nous sommes ensuite focalisés sur l'identification des futurs experts pendant les premiers mois de leur activité au sein de la communauté. Ces experts potentiels sont peu nombreux et leur détection à un stade précoce est très importante car elle garantit la qualité des échanges dans la plateforme sur le long terme. Finalement, nous avons présenté une analyse temporelle des différents usagers dans Stack Overflow. Pour cette étude, nous avons mesuré l'activité des individus pendant plusieurs mois. Au cours du temps, certains utilisateurs restent constants, pendant que d'autres évoluent dans différentes directions (passant d'occasionnel à expert, d'apprenti à expert, etc) dans la plateforme selon leurs capacités et leurs motivations.

2 Introduction

2.1 Problem statement and contributions

Nowadays, with the increasing importance of Social Networks (SN) in our life, we became connected to anyone throughout the world. This is directly related to the emergence of the Web 2.0 during the last decade and its ability to allow people to interact and share knowledge with any one all over the world. With over $3,835$ Billion individuals are connected to internet in 2017[6]. It actually changed the way we seek information. The growing popularity of Social Networks and Question Answering Communities (Q &A C) is the main indicator of how our manner of finding the needed information have became. Communities are built around common topics of interests using social networks, blogs, online communities, etc. These communities interconnect people interested in sharing knowledge and exchanging about their passions or the subjects that they care about. We can find web sites dedicated to several topics such as technologies, religion, cooking, music, traveling, etc. With the development of the computer sciences, Information Technologies (IT) became very popular in both academic and professional areas. People may have to face some difficulties while working with new tools. Several platforms dedicated to this topic have been created and one of the most well known is the Stack Exchange networks, englobing the famous Stack Overflow.

Founded in 2008, Stack Overflow is the largest, most trusted online community for developers to learn, share their knowledge, and build their careers[7]. Over 50 Million of monthly visitors, 7.5 Billion times a developer got helped and an average of 7 visits per user every month. This popularity made Stack Overflow one of the most used question answering community on IT.

This platform is mainly dedicated to allow people to ask question about any issue related to programming. Once the question posted, users provide answers in order to respond helpfully and provide the best content in order to satisfy the question asker. This creates an environment of knowledge exchange.

Lately, Stack Overflow has known a growth of popularity due to number of companies that are using this platform as a hiring tool by publishing job opportunities[8]. The expertise gained in Stack Overflow can be an indicator of how valuable a person is on

[6] http://www.internetworldstats.com/stats.htm
[7] https://stackoverflow.com/company
[8] https://stackoverflow.com/jobs

a given topic. Thus, increasing the popularity gathered in the platform can encourage people to over-top the other members and be spotted out. Therefore, identifying expert users among the high number of persons registered in this platform may be considered as challenging task. Besides the huge amount of job opportunities offered in this community, the fact that one sharing his/her expertise is very helpful to the other members of the community.

As the number of experts is relatively small compared to the population present in these web sites, expertise detection in Question Answering Community (Q&A C) is a very important and challenging task that has to be achieved frequently. Actually, these users are in charge of maintaing of the quality of the exchanges in any Q&A C. Moreover, the rule of 80-20 can be applied to these web sites, where approximately 20% of the users provide 80% of the content according to (Guo et al., 2009), (Matei et al., 2017).

Over time, the behavior of users evolves within a in a Q&A C. Some of them sustain the same level of activity, where other users get very involved within their community. They became more active, engaged and even able to provide answers to the newbies that recently joined the platform. The temporal aspect of users is a very important part of the evolution of the hole community.

In Q&A communities, contributors gain knowledge in some topics by learning from their peers. This analysis will allow us to identify users' activity patterns in Stack Overflow. Besides, the temporal aspect related to Q&A C, the identification of potential experts is very challenging task. These future experts will guarantee the quality of the posts in the communities over time.

The main purpose of community managers is to identify expert users and keep them motivated and active over time. To do so, managers offer them some responsibilities within the community such as the creation of tags, closing questions and moderating exchanges, etc[9].

This may lead us to search for highly expert contributors among millions of users in the community. Looking for them in a traditional way, analyzing their posts and consulting their profiles manually can not be considered owed to the continuous growing number of individuals joining the community. The main goal of this thesis is to propose an automatic model that allows to detect these users. To do so, we have to investigate the data provided by Stack Overflow.

Focusing on this problem related to question answering communities, we summarize the questions that will be treated and answered in this thesis:

- Q_1**: What does the data provided by Stack Overflow will guide us to study the most important features for users?** The platform offers a large

[9]https://stackoverflow.com/help/privileges

amount of data, every user has several features. However, we have to analyze, determine the relation between them and finally identify the most important attributes. The defined attributes will help us to study the performance of users in the community.

- Q_2: **What kind of users do we find in Q&A C?**

  When we are in a well known platform such as Stack Overflow, we may cross several users with different behaviors in the community. They act differently according to their motivations, abilities and their willing to be recognized by their peers.

- Q_3: **What will allow us to distinguish between experts and the other users?**

  The number of experts is small when we see the total number of individuals present in Stack Overflow. In order to detect these "core" contributors among the rest of users, we have to define a measure that will allow us to estimate the degree of their expertise and identify them.

- Q_4: **Can we identify potential experts?**

  Every day, a lot of persons register to the platform. We are interested in a possibility of detecting future experts during the first months of their registration.

- Q_5: **What is the possible evolution of users over time in the community?**

  In Q&A Communities, users evolve differently over a long period of time. The main goal of these users is to help each other, we can find some individuals that may gain knowledge and become very active in order to learn and help their peers. However, not all of them benefit from this source of knowledge. It would be interesting to study the different evolution of different users during their activity in the platform.

Figure 2.1 details the main contributions of this thesis in terms of inputs and results.

In a first place, we focused on the problem of experts detection in Q&A Communities. To do so, we will extract several features characterizing every user in order to provide a general measure of expertise. We will provide an analysis of the data characterizing every user and perform a statistical study according to the importance and the correlation of these features. A user can be described according to the number of questions asked, answers provided and the votes generated by these posts. Any dataset

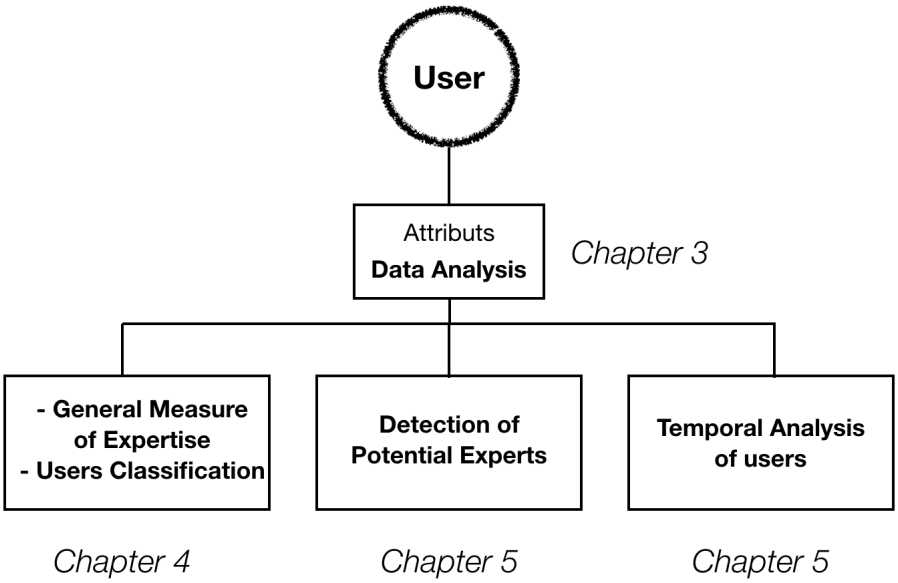


Figure 2.1: Overview of the framework proposed in this thesis to analyze Q&A Communities

from online communities may have imperfect information or not fully available. Yet, to make a decision based on these imperfect data makes the task of identifying experts more challenging. So far, the literature offers tools allowing us to deal with these imperfections. Several theories have been proposed to do so, such as theory of fuzzy sets (Zadeh, 1965), possibility theory (Dubois and Prade, 1988) and the theory of belief functions (Dempster, 1967). The theory of belief functions is a strong tool for representing imperfect information and dealing with uncertainty. Besides, it allows us to manage the combination of pieces of information in our case it will be users attributes.

The combination process will help during the decision making with the classification of a user as an expert or a non-expert.

The theory of belief functions will be the key that will allow us to deal with and manage the uncertainty during the data representation, the combination process and finally for users classification. This general measure of expertise will provide us a general overview on how knowledgeable a user is compared to his peers.

Some users are very passive in the community while others are present to share their expertise and help the community members. Here, with the expertise measure we will define the different roles played by users in the platform. Next, we will focus on detecting potential experts few months after joining the community. As they will be the "core" users of the community, managers have to distinguish them because of their

importance and the responsibilities that they will take on over time.
Later, we focus on providing a temporal analysis on users' behavior. We perform this study by building several time series of number of questions, answers, best answers given by users, and the scores generated for every post. We analyze and study on the evolution of the users expertise over time, their ability, and their motivation to provide helpful answers and share their knowledge over the time series.

2.2 Outline

This thesis is organized in four chapters as follows:

- In Chapter 3, we present a review of the state of art related to question answering communities. To do so, we present some of the most popular survey's related to these communities and the main investigations proposed. Next, we present the very famous web site Stack Overflow. After that, we detail several methods for detection expert users in question answering communities. We divide them on two major methods: ranking-based methods and attributes-based methods. Then, we enumerate different types of imperfect information and how to handle them. One of the proposed methods is to use the theory of belief functions, as the key to deal with uncertainty and combine various types of information. Finally, we present the basic background of this theory illustrated with several examples.

- In Chapter 4, we answer to questions Q_1 and Q_2. So, we present the first contribution which deals with analysis of real data provided by Stack Overflow. In fact, we propose two types of analysis. The first analysis is founded on the Principle Component Analysis (PCA). As every user is characterized by several attributes, we use the PCA because it allows us to reduce the data to its basic components. For the second analysis, we use the mixed classification to determine the number of clusters that may occur in our data. As we are dealing with a huge amount of data, classical techniques can not be applied. To deal with this issue, we use the mixed classification which is the combination of both non hierarchical and hierarchical techniques. Once this step achieved, we conclude that we are facing three main categories of users in question answering community.

- In Chapter 5, we answer to question Q_3. Here, we present our model for characterizing users in Stack Overflow using on the theory of belief functions. We enumerated the hypothesis associated to every class identified in the previous chapter. We select some attributes describing every user. Depending on those attributes, we build our model for the estimation of the general expertise and classify the users. Next we compare our method to the reputation system of Stack Overflow and the Gaussian Mixture Model. We use some internal criteria

for the evaluation of the quality of each clustering. Later, we propose a human evaluation of the three clustering methods. For this step we labeled 500 users and measure some indicators about the efficiency of every method.

- In Chapter 6, we respond to questions Q_4 and Q_5. Thus, we focus on two important issues, the detection of potential experts and a temporal analysis of users in question answering community. For the first part, we compare the results of users few months after they joined the community and the results of the general expertise measure. Our approach founded on the theory of belief functions proposes some interesting results for this task. At the end, we propose a temporal analysis for every class and how can users evolve in the community based on their activity and their motivation.

- Finally, in Chapter 7 conclusions are drawn and some perspectives of this thesis are presented.

Besides these chapters, we have two appendices: Appendix A related to the internal criteria for the evaluation of the quality of the clustering and Appendix B related to the publications resulting from this thesis.

3
Background Review

Knowing ignorance is strength. Ignoring knowledge is sickness
Lao Tse (500B.C.)

Contents

3.1 Introduction

In historical records, it seems that the literature lacks of a detailed definition of term as expert or expertise. Thus, terms such as "masters," "teachers," and "professors" are usually used to describe highly skilled persons, and any reference to the word "expertise" has a general nature.

The earliest recorded educators, including Plato and Socrates, often viewed expertise in what can be described as a "whole man" approach, a holistic view that included aspects of knowledge, skills, and morality to achieve "virtue" in the learner (Voss and Wiley, 2006).

During the late nineteenth century, the study of expertise has began. The primary publications occurred in the late twentieth century with the work on chess in (Groot, 1965), (Chase et al., 1973) and (Chase and Simon, 1973). These researches, have compared expert, middle-range, and novice performance. They have demonstrated the importance of recognizing functionally related "chunks" of chess pieces.

Encyclopedias describe an **"Expert"** as "one who is very skillful and well-informed in some special field" as presented in (Webster, 1968), or "someone widely recognized as a reliable source of knowledge, technique, or skill whose judgment is accorded authority and status by the public or his or her peers."

For (Ericsson, 2006), **"Expertise"** then refers to the characteristics, skills, and knowledge that distinguish experts from novices and less experienced people. In some domains, there are objective criteria for finding experts who are consistently able to exhibit superior performance for representative tasks in a domain. In this thesis report, we embrace the definition presented by (Ericsson, 2006) to describe an expert.

(Forestier et al., 2012) proposed a survey on roles in social networks. If a role is previously defined and its recognition in a network is based on the identification of a number of criteria that can be considered as defined and satisfied by some users. Two major roles have been spotted out in this study: experts and influencers. An Influencer is defined *"as a person who has the ability to influence the decisions or thoughts of other people inside a social network."* These users are more focused on marketing issues. While experts are more present when we are talking about knowledge sharing especially in Question Answering Communities (Q&A C).

Experts detection in Q&A C has received great attention in the literature (Pal et al., 2011), (Pal et al., 2012a), or recently (Srba and Bielikova, 2016b). As we are manipulating real world data, we have to face some imperfections resulting from the information itself or the source of information. The assumption of an uncertain information leads us to deal with a lack of knowledge that we have to consider when we treat any data from online communities. Several theories have been proposed to manage this kind of imperfections such as the theory of belief functions, possibility theory, or fuzzy sets theory.

In section 3.2, we will introduce basic notions related to Question Answering Communities. Besides the most important surveys on these online forums, we will present the very well-known forum Stack Overflow and its reputation system. We will also criticize this flawless measurement and spot out its problems. Next, in section 3.3, we will present some methods proposed in the literature for experts detection by identifying two major approaches: ranking-based approaches and attributes-based approaches. Section, 3.4 recalls the imperfections that occur in online communities and details their typology. In the last section 3.5, we will present the fundamental notions of the theory of belief functions such as mass functions, some particular cases, combination rules and the pignistic transformation used for decision making. We added several examples to illustrate how do we manipulate this theory.

3.2 Question Answering communities

Question Answering Communities have emerged for the last few years. They have become the most used source of information. Well organized, well managed, easy to use, they attracted more and more persons seeking specific information in a given topic. They have established a new paradigm which is learning and collaborating.

(Choi et al., 2012) and (Shah et al., 2014) provided a comprehensive hierarchical classification of online question answering platforms. At the beginning, they differentiated between automatic and human question answering services.

For the first category, automatic question answering sites provide some methods that automatically answer questions asked by humans. In this case, answers do not involve any human interaction.

For the second category, human Q& A web sites, the authors distinguished between expert-based and peer-based systems. On one hand, the expert-based system are managed by small groups of experts rather than an open community as stated by (Choi et al., 2012). Some of these services work on a payment principle referred to as a price-based knowledge market (Chen et al., 2010). Google Answers is an example of these experts based systems (today Google Answers does not accept any new question). On the other hand, the peer-based systems can be split into three groups: community, collaborative and social.

- **Community**: they are pointed out as knowledge exchange communities by (Adamic et al., 2008). They are composed of three major elements. First, they allows information seekers to ask question. Second, contributors submit their answers. Finally, an entire community is build around these exchanges. Quora, Stack Overflow are two of the most well known communities.

- **Collaborative**: they are defined by the same mechanism as the community systems. However, these websites rely on one or few knowledgeable individuals to

Figure 3.1: Stack Exchange

provide answers. Wiki answers is one of the most popular collaborative services.

- **Social**: they use of features from social networking sites as a mean for knowledge sharing. They allow users to ask questions to friends or acquaintances inside their network. They differ from the other types by the fact that the user most likely does not always trust the information source, since it is someone from his/ her personal network (Morris et al., 2010). Facebook, Twitter and Instagram are the most famous social networks nowadays.

One of the most popular platforms is StackExchange[10] as illustrated in Figure 3.1[11]. It is a network of over 150 communities[12] each specialized in a specific topic of interest (technology, business, art, science, etc) including the very famous website Stack Overflow (SO). Several other websites such as Yahoo!Answers[13], Quora[14] are very well known but not as popular as Stack Overflow especially for the topics related to computer science.

A classical question answering process is composed of several steps. Starting with an unsuccessful research for an answer in the traditional information systems, a person will visit a question answering platform. Then, he /she will post a question by providing a

[10] https://stackexchange.com
[11] https://stackexchange.com/about
[12] //stackexchange.com/about
[13] https://answers.yahoo.com/
[14] https://www.quora.com/

title and a description of the encountered issue. The post must be written in a natural language, annotated with keywords. The question must be understandable and precise so the asker can receive adequate answers. Once this step achieved, the platform will propose the question to potential answerers who are willing to share their knowledge and provide high quality answers. It is a major aspect of the process because if the question is not proposed to suitable users, it is very likely to stay unanswered through time. In addition, users can comment, edit and vote for both questions and answers. Next, the question asker can choose the best answer among all of the provided ones. Finally, the question is marked as resolved and archived in the platform.

Figure 3.2 shows how a user behaves in question answering platform. A user has two types of features: community and outside community features. For the Q&A C features we can distinguish between activity, expertise, temporal features and popularity.

- **Activity**: describes what a user does within the community: number of questions, answers, best answers,
- **Expertise**: measures the degree of knowledge of a user,
- **Temporal**: measures the time activity since the user joined the community,
- **Popularity**: estimates how popular the user is among the other members of the community.

The outside Q&A C features we can distinguish between a user's activity outside the community, and the information posted when he/she filled his profile while joining the community.

A user can post both questions and answers. A question is characterized by its length (it must not be very long, otherwise people will not take time to read it and respond), its readability (it has to be written in a natural language, precise, easy to read and understandable to other users), the keywords (to describe the issue) and a specific topic of interest. A question can have none or multiple answers. An answer must have high quality content and be helpful, responding to the asker's needs and expectations.

3.2.1 Surveys on Question Answering Communities

Despite the great number of recent publications focusing in question answering communities, the first researches started only on the mid 90's. The first survey was presented by (Ackerman and McDonald, 1996). However, the first complete survey on this field was presented in late 2000 by (Chirag Shah and Oh, 2009). This is directly related to the recent emergence of Web 2.0. where online forums, and platforms have gained more and more popularity. In Table 3.1 we present a small resume of the best surveys on question answering communities.

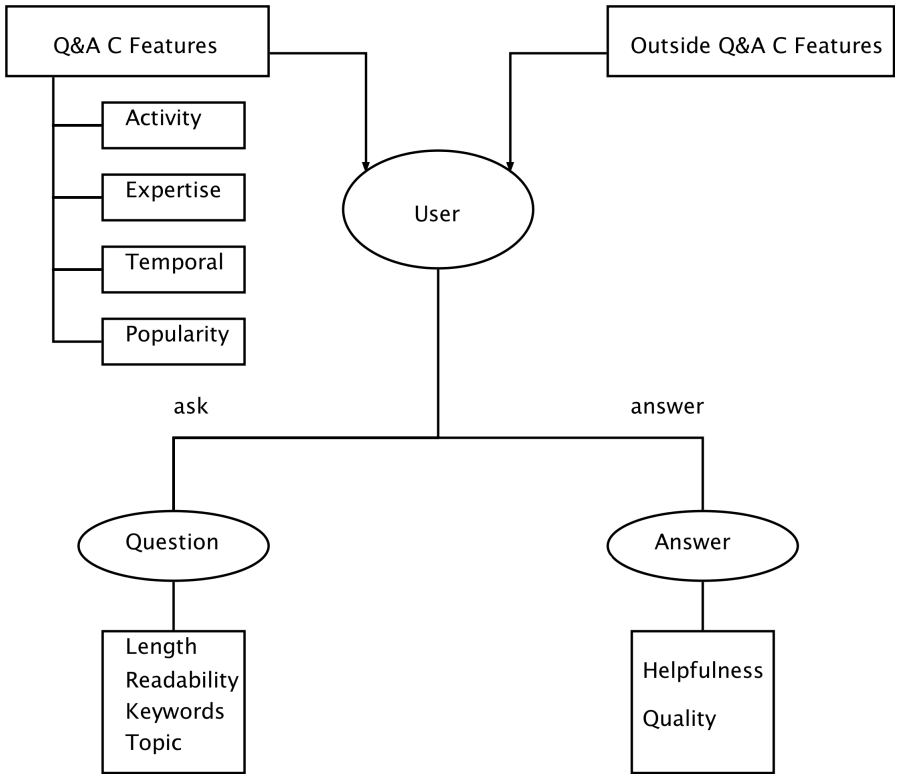


Figure 3.2: Users in Question Answering Communities (inspired by (Srba and Bielikova, 2016a))

In (Chirag Shah and Oh, 2009), the authors treated reduced number of studies due to the recent popularity of online platforms. The authors distinguished two major approaches: the content and users based researches. On one hand, the content-based studies target the evaluation of the quality of the answers. On the second hand users-based studies were more diverse. They focused on the roles played by users during the question answering process and the detection of authoritative users. Few years later, in (Gazan, 2011) investigated collaborative question answering platforms. They identified four approaches related to these web sites:

- Classification and retrieval of questions,
- Classification and evaluation of the quality of answers,
- Satisfaction of users in these communities,
- Evaluation of the motivation, reputation and how users perceive experts.

The authors also proposed valuable ideas for future research.

In (Furlan et al., 2013), the authors have privileged the study of the question routing approaches. They covered three basic processing stages related to the three major problems of question answering system implementation: question analysis, question forwarding, and users' knowledge profiling.

The question analysis part deals with question processing, question forwarding treats the matching, ranking forwarding questions. Finally, user profiling focuses on some internal (user activities in the platform) and external criteria (other social networks, blogs, email). This study presented a full analysis on question routing in online communities.

In (Srba and Bielikova, 2016a), the authors presented the most complete survey published that covers 265 articles published between 2005 and 2014. They come up with a classification of the different approaches based on the problems treated in every one. They propose a very convenient study for any person starting in the research area of Question Answering Communities.

(Tuna et al., 2016) proposes an analysis of users' characterization in social networks. They provided an overview of users' attributes selection such as gender, age, location, etc. They also identified how a user can behave on social networks as deceptive behavior (users providing false information), privacy behavior (how to measure and evaluate the privacy), radicalism and reactions to attacks. They analyzed how a user reacts when he/she is under attack from peers. By attacks authors focus only on hacks.

Table 3.1: Description of Approaches on Question Answering Community

(Chirag Shah and Oh, 2009)	Content-centered approach Users-centered approach
(Gazan, 2011)	Question classification and retrieval Answer classification and quality evaluation Users satisfaction Motivation, reputation, perceived authorities
(Furlan et al., 2013)	Question routing methods (recommendation of posts to potential answerers)
(Srba and Bielikova, 2016a)	Exploratory study on QAC Content and users modeling Adaptive support
(Tuna et al., 2016)	Attributes selection User's behavior Identification of spammers

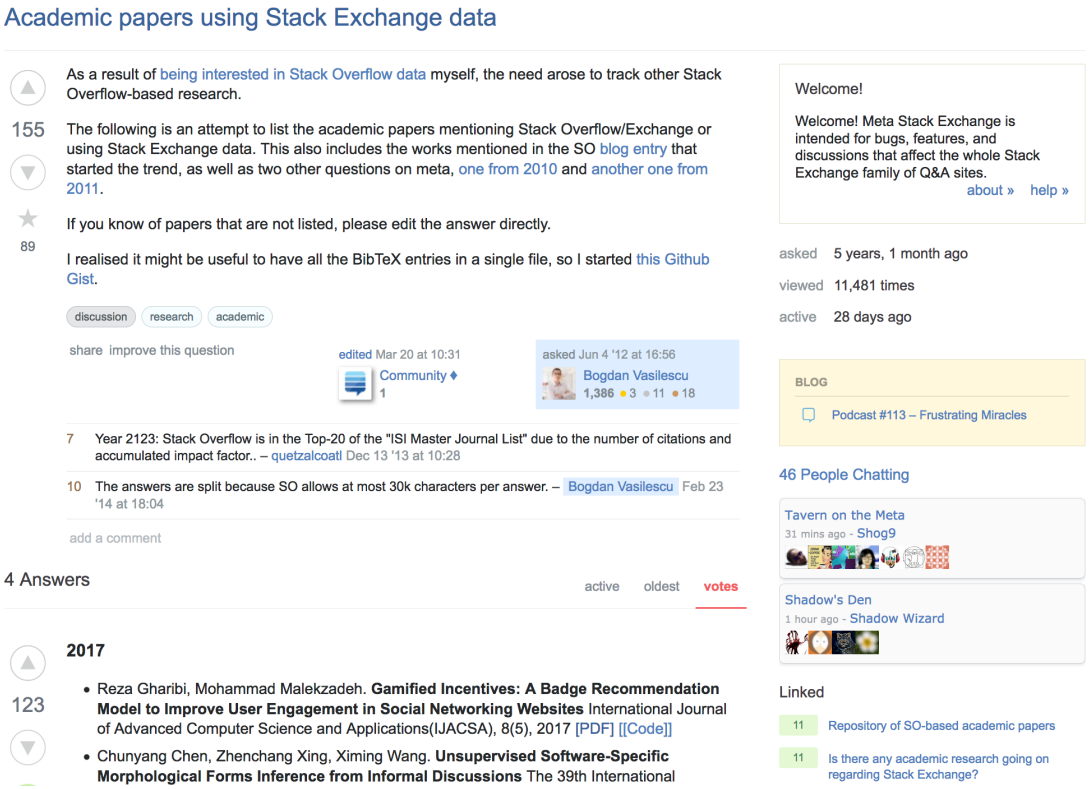


Figure 3.3: Example of a question and answers in Stack Overflow

3.2.2 Stack Overflow

One of the most popular platforms is Stack Overflow (SO)[15]. It is the largest online community for programmers. Created in 2008 by Jeff Atwood and Joel Spolsky. The name for the website was chosen by voting in April 2008 by readers of Coding Horror, Atwood's popular programming blog. The name was chosen because of its meaning. A **"stack overflow"** is an undesirable condition in which a particular computer program tries to use more memory space than the call stack has available. In programming, the call stack is a buffer that stores requests that need to be handled[16].

Here, users can post questions, answers them, vote positively or negatively for both answers and questions in order to express their opinion on the quality of the posts as shown in Figure 3.3.

In Stack Overflow, users can ask questions about[17]:

[15] http://stackoverflow.com
[16] http://whatis.techtarget.com/definition/stack-overflow
[17] https://stackoverflow.com/tour

- Specific programming problems.
- Software algorithms
- Coding techniques
- Software development tools

Reputation

Stack Overflow proposes a reputation system to reward active users. Actually, reputation[18] is the summary of users' activity on the web site. All users start with one reputation point, and reputation can never drop below 1. It is earned by convincing other users that he/she knows what he/she is talking about. Indeed, reputation reflects how involved a user is in the community and how other people see him/her. If this value is high, it means that a user is able to post fair questions or/and answers and how well he/she can communicate and interact with his/her peers. It also means that we can be in presence of a knowledgeable person. However, we assume this measurement as flawed.

Most of the users of Stack Overflow aim to win as much reputation points as possible in order to obtain privileges like creating tags, moderating the forum etc. The reputation is defined according to the system presented in Table 3.2.

Table 3.2: Gratification system of Stack Overflow

Action	**Reputation**
Answer voted up	+10
Question voted up	+5
Accepted answer	+15 (+2 to question asker)
Question voted down	−2
Answer voted down	−2 (−1 to voter)
Spammed answers	−100
Accepted answer to bounty	+bounty
Offer bounty on question	-bounty

Every post (can be rather a question or an answer) can be submitted either to positive or negative votes. A positive vote is a reward for the author, while the negative one penalizes him. Each person who posts a question is allowed to choose the best answer that seems to be the most helpful allowing his/her owner to gain reputation points.

[18] http://stackoverflow.com/help/whats-reputation

A bounty[19] is a special reputation award given to answers. It is supported by the reputation points of a user who want to offer them as a reward for a satisfactory answer. A bounty may help attract more attention and more answers.

Reputation actually reflects how a user is involved in the community and how other people see him. Most of the users within this platform aim to win as much reputation points as possible in order to obtain privileges like creating tags, moderating the forum etc. Any gamification system is proposed to improve and encourage user's engagement, productivity, learning process, and evaluation as defined by (Huotari and Hamari, 2012). Moreover, for some companies proposing job offers in the platform the reputation can be an indicator on the user's ability to provide answers and works with a programming tool.
However, this measurement does not take into account the quality of the posts. Indeed, if a question is considered as simple in a very popular topic, answers will be numerous and quick, creating a competitive spirit within the community. Therefore, if a question is seen as difficult in a less popular topic, contributors may not take the risk to post answers, due to their lack of knowledge or the risk of being evaluated negatively by the community and of losing some reputation points.

3.3 Expertise detection in Question Answering Communities

The success known of Question Answering Communities have influenced several academic research interests. Many studies have investigated users' motivation and interest for participating in these platforms.

User expertise is closely associated with several different terms such as user authority, user reputation. The common characteristic of all these terms is that they refer to a user-related measure that captures an amount of user knowledge and his/her potential to provide high-quality content.

Expert finding addresses the task of identifying the right person with the specific skills and knowledge to solve a problem (Balog and de Rijke, 2009).

In order to identify experts in online question answering communities we can distinguish between two major approaches: ranking and attribute based approaches as described in (Bouguessa and Romdhane, 2015), (Sahu et al., 2016). While the ranking-based approaches aim to calculate a metric or a score per user which is used to select a specific number of users' described as expert or authoritative. The attribute-based approaches aims to identify a number of features relative to each user. These attributes allow to classify users as expert or non-expert based on machine learning techniques. Both approaches rely on prior knowledge about the manipulated dataset.

[19] https://stackoverflow.com/help/bounty

3.3.1 Ranking-based approach

The ranking based approaches intent to measure a score per user then select the top users as defined by (Tang and Yang, 2012).

(Zhang et al., 2007) introduced the ExpertiseRank which is an extension of PageRank (Page et al., 1999) allowing to compute the expertise score of a users in a question answering community. Besides, the graphical features, this algorithm also includes a metric called "Z-Score" based on both the number of answers and the number of questions asked by a given user. Their result supposed that a metric like "Z -Score" outperforms over complex graph based algorithm such as PageRank. For the latter algorithm, it is very greedy and expensive in a computational axis. The main issue with the Z-Score method is that it takes into consideration only few indicators about users.
(Ramage et al., 2009) proposed a topic model that constrains Latent Dirichlet Allocation (LDA) by defining a one-to-one correspondence between LDA's latent topics and user tags. This allows Labeled LDA to select the user's topical interests based on their former answers. At the same time, the expertise level is measured using a collaborative voting mechanism. The problem with an LDA based technique like this method is that the correlation between labels can not be taken into consideration.

(Kao et al., 2010) proposed a hybrid approach for experts identification in online communities. They proposed to combine user knowledge profiles, user reputations and link analysis to find authoritative contributors for a given category of questions. This method is based on the reputation which we consider as a flawless measure, and it is an extension of the PageRank meaning that is is a greedy algorithm.
(Yang et al., 2013) introduced the CQARank algorithm that measures user interests and expertise score under different topics. Their proposal is based on hybrid generative model with Gaussian Mixture Model (GMM) and outperformed several approaches such as PageRank.

For (Movshovitz-Attias et al., 2013), the authors proposed an analysis of Stack Overflow's reputation system. They focused on the contributors participation model. They considered the reputation as measurement of expertise. Any user with a reputation grater than 2400 points is an expert. However, their approach seems to be strict because it is only based on the value of the reputation gathered during users activity on the platform.
(Song et al., 2013) attempted to discover leading users on Quora. Authors introduced a leading capacity model, which considered three user characteristics: authority, activity, and influence. This approach have been tested on only one topic. Authors select the number of top users without any classification method made in order to comfort the results.

Another approach is proposed in (Yang et al., 2014) that is not founded on the reputation measure. They defined a metric called "Mean Expertise Contribution" (MEC)

that takes into account two indices: the debate generated by a question and the utility of the provided answers. The first index is related to the number of answers proposed for a given question. The second one is calculated according to the rank of an answer among all the answers provided. The authors tested they model on only one single and very popular topic. It has to be tested on other topics less famous in order to see the effectiveness of this method without as much participation by users.

The main issue with all the ranking-based-approaches, is that the number expert users is defined in at first for every method. In Question answering communities, we can not predefine from the beginning the number of contributors we consider as expert. Moreover, some of the methods described below such as ignore an important information which is the number of best answers provided by a user for the measurement of their scores.

3.3.2 Feature-based approach

Attributes based approaches aim to identify a number of features for the users and then apply machine learning techniques in order to classify users.
In (Bouguessa et al., 2008), authors used a Beta Mixture Model (BMM) to identify authorities in online communities. They rated users in Yahoo!Answers based on their activities. A BMM is a statistical distribution has a support range of $[0, 1]$ as defined by (Ma and Leijon, 2011). It is applied to model events that take place in a limited interval and is widely used in financial model building, and social networks. The main issue with this work is that the authors choose the number of individuals that are chosen as experts.

(Pal and Counts, 2011), authors focused on a number of attributes in Twitter users in order to identify the most authoritative persons. They performed a clustering based on a Gaussian Mixture Model (GMM) in order to separate users between influential and non-influential. After that they performed a ranking mechanism in order to select a specific number of the most authoritative users. This metric is based on user activity, therefore partially convenient for Stack Overflow because it an approach based on reputation .

(Chan et al., 2010) and (White et al., 2012) selected a set of user features (number of users they answered to, number of users answering their questions, the mean and standard deviation of posts per topic, etc) and performed a clustering over them. They examined the resulting clusters and manually fused users having similar behavior within the community. They only used hard clustering techniques to define a high number of potential classes of users that might be readjusted.

(Pal et al., 2012b) examined the selection of question preferences allowing them to identify several characteristics for the community users. These characteristics are used as features to classify users between experts or non-expert. The authors developed a probabilistic model supported by a machine learning algorithm to detect experts

and identify potential experts in online question answering communities. The main limitation with this method is that it is based on a supervised machine learning, the resulting classification is highly dependent on the training dataset.

(Furtado et al., 2013) performed a clustering on user's attributes (number of questions, number of days a user was active) to classify a random set of users in groups having similar contributions behavior. Ten profiles were identified and manually labeled, for example, occasional, unskilled or expert answerer, answer activist, etc. This method generated ten classes, which can be very greedy and hard to represent for comparative models.

In (Van Dijk et al., 2015), authors proposed an early detection of topical expertise based on the attributes of users and mainly the number of accepted answers. They also based their approach using textual, behavioral and temporal characteristics of the users.

Recently in (Sahu et al., 2016), authors extended the work presented in (Bouguessa and Romdhane, 2015) called Multivariate Beta Mixture Model (MBMM). Their model based on a Beta Mixture Model (BMM) evaluates the profiles of users to distinguish between authoritative and non-authoritative users based on voting mechanism. Their challenge was to locate important users by mining textual and meta data features. The BMM is used as an unsupervised clustering approach.

The main advantage of these methods is that the results are easy to interpret since each clustering result is attached to a set of different values in characterizing the used features. However, none of the approaches cited below use uncertainty theory of any combination process in order to classify users in Q&A Communities.

3.3.3 Other methods

In (Liu et al., 2011) and (Aslay et al., 2013) proposed to create competition-based expertise networks, which combine all available information into one community expertise network. Their main approach is defined according to two hypothesis. The first hypothesis concerns the fact that answerers providing best answers have higher expertise compared to the others. The second one considers the assumption that a person who provides answers has more expertise than the question asker. (Liu et al., 2011) applied this network to estimate the expertise of users by applying competition-based model. Experiments on the real world data obtained from Yahoo! Answers confirmed that the competition-based models are able to significantly outperform standard graph-based baseline methods, such as HITS and PageRank.

3.3.4 Summary of expert detection

Most of the approaches proposed in the literature are based on the content of the posts and text mining techniques. None of the studies mentioned have used the theory of

uncertainties to manage the data imperfections. In Chapter 5 we propose a measure of expertise based on the theory of belief functions. This measure will allow us to detect experts in Q&A C.

Table 3.3 summarizes some of the most used approaches in the literature for experts detection. Here, we present the name of the method, if the authors propose a general or topical measure of expertise. Is it based on ranking or features based methods. We also indicate if there was a textual analysis of the data and the platform used for the experiments.

Table 3.3: Summery of expert detection in Q&A C

Authors	**Method**	**General**	**Topical**	**Ranking**	**Feature**	**Textual**	**DataSet**
(Zhang et al., 2007)	Z-Score		•				Java Forum
(Zhang et al., 2007)	ExpertRank		•		•		Java Forum
(Ramage et al., 2009)	Labeled LDA		•	•		•	Yahoo
(Bouguessa et al., 2008)			•	•		•	Yahoo
(Song et al., 2013)			•	•		•	Quora
(Movshovitz-Attias et al., 2013)		•			•		Stack Overflow
(Yang et al., 2014)	MEC		•		•		Stack Overflow
(Pal et al., 2012b)		•			•		Stack Overflow
(Van Dijk et al., 2015)			•		•	•	Stack Overflow
(Sahu et al., 2016)	MBMM		•		•	•	Stack Overflow AskUbunto

3.4 Uncertainty in Question Answering communities

(Smets, 1996) defines uncertainty as *"partial knowledge of the true value of the data. It results in ignorance (etymologically not knowing). It is essentially, if not always, an epistemic property induced by a lack of information. A major cause of uncertainty is imprecision in the data".*
In (Kasneci et al., 2011), the authors identified three levels of uncertainty in question answering communities. The first level is related to the extraction and integration of the data. The second one deals with information sources, meaning the users of these platforms. The third level covers the uncertainty of the information itself. In the considered case, we are more interested in the evaluation of the sources and the part of uncertainty related to them.
The main issue in these communities is that we are dealing with users that we do not usually have an *a priori* knowledge about them. We ignore everything about the sources' reliability, or expertise. In order to deal with uncertainty, several theories were proposed such as probability theory (Reyni, 1962), possibility theory (Dubois and Prade, 2015) and the theory of belief functions (Dempster, 1967). The latter can be presented as a generalization of the other theories. Besides it offers a rich tool able to manage different types of data imperfections.

When manipulating uncertainty, information fusion can be an interesting solution to obtain relevant information. Data fusion based on the theory of belief functions has been widely used in classification, image processing (Khaleghi et al., 2013), etc. and in social networks (Attiaoui et al., 2015), (Nguyen and Huynh, 2016), (Dlala et al., 2015), and more recently (Zhou et al., 2016) and (Jendoubi et al., 2017).

We will use the mathematical background provided by the theory of belief functions. This will help us to consider the problem of early identification of potential experts with an uncertain point of view.

Among the reasons justifying the use of the theory of belief functions we can cite:

- The possibility to represent ignorance and all kind of imperfect information.
- The mathematical representation allow us to model several types of information through a rich modeling framework.
- The robust combination rules present in the framework.
- The management of conflict: actually during the information fusion, some conflict between the masses occurs. Thus, the Dempster's combination rule measures the conflict and redistribute it.
- The combination process helps for the decision making.

When we are in presence of pieces of information, characterizing the real world, we are sure that they are imperfect. Here, we introduce three main types of imperfect information.

3.4.1 Imprecise information

It is characterized by the information content. In other words, it is related to the information itself. Let's consider, for example, the age of a person. We say "John's age is between 25 and 30".
Thereby, we formally describe it as: $age(John) \in \{25, 26, 27, 28, 29, 30\}$. It represents an imprecise information. In this case, we are unable to determine the exact age of John.
Thus, when we deal with this kind of information, we are not able to fully grasp the real world situation.

3.4.2 Uncertain information

It is the result of a lack of information about the real world. It is related to the source providing the information.
Any uncertain information has an uncertain "score" which can on one hand be numeric (The probability that Donald Trump will be relected in 2020 is 0.1), and on the other hand symbolic or linguistic (I believe that John is 29 years old).

3.4.3 Inconsistent information

Where no value is compatible with the information. For example "I am a PhD Student in Information Technology", and "I am a Doctor in Information Technology". I can not be a doctor and PhD Student at the same time. The conflict between the two pieces of information can lead us to an inconsistent deduction.

3.5 Basics on the theory of belief functions

The theory of belief functions started with the work of A. Dempster (Dempster, 1967). The aim of his researches was to model mathematically information that can not be described by a precise probabilistic distribution. To do so, he developed the notions of the lower and upper probabilities framing the exact distribution. Using that, he was able to represent more precisely the observed data.

Later, in his book "A mathematical Theory of evidence", Shafer presented the information defined by an expert, where basic belief assignments have two functions: a credibility and a plausibility function corresponding respectively to the lower and upper

probabilities of Dempster.

The theory was further developed by (Smets and Kennes, 1994), who proposed the Transferable Belief Model (TBM). This model presents a pignistic probability induced by a belief function which is built by defining a uniform probability from each positive mass. Moreover, in terms of upper and lower probabilities, it can be considered as the center of gravity of the set of probabilities dominating the belief functions. He also introduced new tools for information fusion and decision making.
The objective of the theory of belief functions is to represent information transmitted by a source concerning an event. A belief function must take in consideration all the possible events on which a source can describe a belief. Based on that, we can define the frame of discernment

3.5.1 Frame of discernment

The frame of discernment is a finite set of disjoint elements noted Ω. It is defined as $\Omega = \{\omega_1, ..., \omega_n\}$. This theory allows us to affect a mass on a set of hypotheses not only a singleton like in the probabilistic theory. Thus, we are able to represent ignorance, imprecision, etc.

3.5.2 Basic belief assignment

A *bba* is defined on the set of all subsets of Ω, named power set and noted 2^{Ω}. It affects a real value from $[0, 1]$ to every subset of 2^{Ω} reflecting sources amount of belief on this subset. A *bba* m verifies:

$$\sum_{X \subseteq \Omega} m(X) = 1. \tag{3.1}$$

Example 1 *Let us consider a question posted by a user u_1 in the online community. Two other users u_2 and u_3 will read the question and will try to identify the profile of the author u_1 : is he an expert or not.*
Thus, the frame of discernment Ω is formed of Expert (E) and Non Expert (NE): $\Omega = \{E, NE\}$. The corresponding power set is $2^{\Omega} = \{\emptyset, E, NE, E \cup NE\}$.
Along this section, this example will be used to illustrate some basic notions in the theory of belief functions.

A very common assumption advocates the existence of a closed world. In other words, all the possibilities are represented in Ω, and defined as:

$$m(\emptyset) = 0. \tag{3.2}$$

On the contrary, if we accept the case that, there exists any other possibilities that is unrepresented in Ω, we have:

$$m(\emptyset) > 0. \tag{3.3}$$

Here, Smets supposed that we are dealing with information in an open world, which means that decisions are not exhaustive. This was introduced by Smets.

3.5.3 Focal elements

Considering a set A in 2^{Ω} is a focal element with a mass m if an elementary mass is positive $m(A) > 0$. The set of focal elements of m is noted $\mathbb{F}(m)$.

Example 2 *Let us take the same example of evaluation of the author of the question. To express their beliefs on the question asker, the belief holder, which is the user* u_2 *will say that this person is an expert at* 80% *and* 20% *ignorance (*u_2 *does not know). User* u_3 *would say this person could be an expert with a belief of* 70% *and* 30% *of ignorance. We obtain the following mass functions:*

$$m_{u2}(E) = 0.8, \quad m_{u2}(\Omega) = 0.2 \tag{3.4}$$

$$m_{u3}(NE) = 0.7, \quad m_{u3}(\Omega) = 0.3 \tag{3.5}$$

3.5.4 Particular belief functions

Mass function is the common representation of evidential knowledge. Basic belief assignments are degrees of support justified by available evidences. This section recalls some particular mass functions.

Categorical mass functions

A categorical mass function is a normalized mass function which has an unique focal element X^*. This mass function is noted $m(X)$ and defined as follows:

$$m_{X^*}(X) = \begin{cases} 1 & \text{if } X = X^* \subset \Omega \\ 0 & \forall\ X \subseteq\ \Omega \text{ and } X \neq X^* \end{cases} \tag{3.6}$$

We distinguish two particular cases of categorical mass functions: the vacuous mass functions when $X^* = \Omega$ and the contradictory mass functions if $X^* = \emptyset$.

Example 3 *Assume that the user u_2 identified user u_1 as an Expert, $X = E$. The corresponding mass function is a categorical belief function expressed as:*

$$m(E) = 1.$$

Vacuous mass functions

A vacuous mass function is a particular categorical mass function focused on Ω. It means that a vacuous mass function is normalized and has an unique focal element which is Ω. This type of mass functions is defined as follows:

$$m_\Omega(X) = \begin{cases} 1 & \text{if } X = \Omega \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

Vacuous mass function emphasizes the case of total ignorance.

Example 4 *The user u_3 supposed that he identified that u_1 is either an expert or not, where $X = E \cup NE$. The corresponding mass function is a vacuous belief function expressed as : $m(X) = 1$.*

Simple support mass functions

Simple support mass functions, introduced in (Smets, 1995), are a special type that allow us to model both of the uncertainty and imprecision according the following equation:

$$\begin{cases} m(X) &= 1 - \omega,\ X \subset \Omega \\ m(\Omega) &= \omega \\ m(Y) &= 0,\ Y \neq X \subset \Omega \end{cases} \tag{3.8}$$

where the mass on $m(\Omega)$ represents the ignorance.

Example 5 *Suppose the frame of discernment $\Omega = \{E, NE\}$. We assume that a simple support mass function for the same example expressed by the following equation:*

$$m_{u2}(E) = 0.6, \quad m_{u2}(NE) = 0.2, \quad m_{u2}(E \cup NE) = 0.2$$

m a simple support function focused on E.

Dogmatic mass functions

A dogmatic mass function is a mass function where Ω is not a focal element. A dogmatic mass function is defined as follows:

$$m(\Omega) = 0. \tag{3.9}$$

Bayesian mass functions

A Bayesian mass function is a mass function which all focal elements are elementary hypotheses. It is defined as follows

$$\begin{cases} m(X) \in]0,1] & \text{if } |X| = 1 \\ m(X) = 0 & \text{otherwise} \end{cases} \tag{3.10}$$

As all focal elements are single points, this mass function is a *probability distribution.*

Example 6 *Suppose the user u_3 assigns the following mass functions on u_1:*

$$m_{u3}(E) = 0.8\ , \qquad m_{u3}(NE) = 0.2$$

This mass function is Bayesian and the corresponding probability distribution is

$$p(E) = 0.8\ , \qquad p(NE) = 0.2$$

Consonant mass functions

A consonant mass function is a mass function which focal elements are nested, this mass function is defined by the following:

$$(X_1 \subseteq X_2 \subseteq ... \subseteq \Omega) \tag{3.11}$$

Certain mass functions

A certain mass function is a categorical mass function (a mass supporting an unique focal element) such that its focal element is an elementary hypothesis. This mass function emphasizes the case of total certainty as the source supports only one hypothesis with certainty.

$$m(X) = \begin{cases} 1 & \text{if } X = \omega \in \Omega \\ 0 & \forall X \subseteq \Omega \text{ and} X \neq \omega. \end{cases} \tag{3.12}$$

Example 7 *Suppose the user u_3 is certain about u_1 being an expert.*

$$m_{u3}(E) = 1$$

3.5.5 Combination rules

Belief functions are used to represent information provided by different sources, it is natural than to combine them for an issue of decision making.
Many combination rules have been proposed taking in consideration the nature of the sources.

Dempster's combination rule

The first one was proposed by Dempster in 1967 which is a conjunctive normalized combination rule also called *the orthogonal sum.*
Given two mass functions m_1 and m_2, for all $X \in 2^\Omega$, $X \neq \emptyset$ the Dempster's rule is defined by:

$$m_D(X) = \frac{1}{1-k} \sum_{Y_1 \cup Y_2 = X} m_1(Y_1) m_2(Y_2) \tag{3.13}$$

where $k = \sum_{Y_1 \cap Y_2 = \emptyset} m_1(Y_1) m_2(Y_2)$ is the inconsistency of the fusion (or of the combination) can also be called the conflict or *global conflict.*
The value $1 - k$ is the normalization factor of the combination in a closed world.

The conjunctive combination rule

In order to consider the issues of the open world, like introduced in (Smets 1990), where the author proposed the conjunctive combination rule. Considering two mass functions m_1 and m_2, for all $X \in 2^\Theta$ defined by:

$$m_{conj}(X) = \sum_{Y_1 \cap Y_2 = X} m_1(Y_1) m_2(Y_2) \tag{3.14}$$

We can note $m_{conj} = m_1 \oplus m_2$, and consider $k = m_{conj}(\emptyset)$ as an unexpected solution.

The operator $\oplus$ is associative and commutative but not idempotent.

This combination rule can be extended to N mass functions m_i. We obtain $\oplus_{i\in[1,N]}m_i$, for all $X \subseteq \Omega$:

$$\oplus_{i\in[1,N]}m(X) = \sum_{Y_1\cap\ldots\cap Y_N=X} \prod_{i\in[1,N]} m_i(Y_i) \tag{3.15}$$

The disjunctive combination rule

First introduced by (Dubois and Prade 1986), then by (Smets 1993), we can consider two basic belief assignments m_1 and m_2, after proceeding to a disjunctive combination expressed like follows for all $X \subseteq \Omega$:

$$m_{disj}(X) = \sum_{Y_1\cup Y_2=X} m_1(Y_1)m_2(Y_2) \tag{3.16}$$

The disjunctive combination rule can be used when one of the sources is reliable or when we have no knowledge about their reliability.

Example 8 *Now, we will proceed to the combination of the assumptions made by users u_2 and u_3 about the expertise of user u_1. We keep the values of the masses presented in Example 2. The masses obtained after using the three combination rules defined previously are the following:*
Dempster's combination rule:

$$\begin{aligned} m_D(\emptyset) = 0, \qquad & m_D(E) = 0.5455, \\ m_D(NE) = 0.3182, \qquad & m_D(\Omega) = 0.1364 \end{aligned}$$

Conjunctive combination rule:

$$\begin{aligned} m_{Conj}(\emptyset) = 0.56, \qquad & m_{Conj}(E) = 0.24, \\ m_{Conj}(NE) = 0.14, \qquad & m_{Conj}(\Omega) = 0.06 \end{aligned}$$

Disjunctive combination rule:

$$\begin{aligned} m_{Disj}(\emptyset) = 0, \qquad & m_{Disj}(E) = 0, \\ m_{Disj}(NE) = 0, \qquad & m_{Disj}(\Omega) = 1 \end{aligned}$$

In the literature, we find many other combination rules such as the Proportional Conflict Redistribution $PCR6$ proposed by (Martin and Osswald, 2008) and (Martin and Osswald, 2007) proposed an evolution by defining the PCR 6 or even in (Martin, 2009). Latley, (Chebbah et al., 2015) proposed an other rule allowing the combination of partially independent belief functions.

3.5.6 Discounting

In the belief function framework, knowledge about the reliability of a source of information (or user) is achieved by the discounting operation, which transforms each belief function provided by a source into a weaker, less informative one. The discounting operation is controlled by a discount rate in taking values in $[0, 1]$: if $\alpha = 0$, the belief function is unchanged; if $\alpha = 1$, the belief function is transformed into the vacuous belief function. This transformation means that the information provided by the sensor is completely discarded.
(Smets, 1993) shows that the discounting operation is not *adhoc*, it can be derived from a simple model of sensor reliability. In this model, the sensor can be in two states: reliable or not. In the first case, when we know that the sensor is reliable, the belief function which is provided is accepted without any modification, otherwise when we know that it is not reliable, the information is considered as irrelevant.

$$\begin{cases} {}^{\alpha}m\,(X) = & \alpha m(X) & \forall, X \subset 2^{\Omega} \setminus \Omega \\ {}^{\alpha}m(\Omega) = & 1 - \alpha(1 - m(\Omega)) \end{cases} \tag{3.17}$$

Example 9 *Suppose that the degree of reliability of user u_2 and user u_3 are respectively* 0.3 *and* 0.7. *The discounted mass functions are:*

$$\begin{array}{llll} m_{u2}(\emptyset) = 0 & m_{u2}(E) = 0.24 & m_{u2}(NE) = 0 & m_{u2}(\Omega) = 0.76 \\ m_{u3}(\emptyset) = 0 & m_{u3}(E) = 0 & m_{u3}(NE) = 0.49 & m_{u3}(\Omega) = 0.51 \end{array}$$

3.5.7 Decision making

In (Smets, 2005), the author introduced the Transferable Belief Model (TBM). He afforded new method to to represent, manipulate and combine information from different sources. The TBM is based on two levels:

- **The credal level** (*in Latin credo: I belief*), where beliefs are studied and combined using belief functions in order to preserve as much information as possible during the combination aiming at decision making.
- **The pignistic level** (*in latin pignus: a bet*), where beliefs are used to make decisions and represented by probability functions called the **pignistic probabilities**

When choosing the maximum of the credibility might be pessimistic, on the other hand decision made with the maximum of plausibility can be considered too optimistic. Based on that, the best solution is to use the **pignistic probability** choose the credal level, where for all $X \in 2^{\Omega}$, with $X \neq 0$:

$$BetP(X) = \sum_{Y \in 2^{\Omega}, Y \neq 0} \frac{\mid X \cap Y \mid m(Y)}{\mid Y \mid 1 - m(\emptyset)}. \tag{3.18}$$

Example 10 *Suppose that the degree of reliability of user u_2 and user u_3 are respectively* 0.3 *and* 0.7. *The discounted mass functions are:*

$$BetP_{u1}(E) = 0.6136 \qquad BetP_{u1}(NE) = 0.3864$$

Then user u_1 is an Expert with a pignistic probability of 0.6136.

(Essaid et al., 2014) proposed a method for decision making in the credal level. This rule uses a distance. This decision is made on the most reasonable hypothesis based on the measurement of the distance between a combined *bba* and a categorical *bba*.

3.6 Conclusion

In this chapter, we give an overview of the state of the art of some expertise detection in question answering communities. In the first place, we review on information how experts are identified. We classify them into two main categories, in the first one is link-based approaches and the second one is attribute-based approach. In a second axis, we presented an overview on Stack Overflow and how they proposed a controversial system of expertise identification based on the reputation. Next, we detailed the imperfections that may occur in these online communities. When dealing with real world data we have to manage the uncertainty or imprecision by using the adequate tools such as the theory of belief functions, theory of probabilities, possibilities, etc. In the sequel of this report, imperfect data is modeled with the theory of belief functions. We have introduced the necessary theoretical background, by showing its strength in modeling imprecise, uncertain or incomplete information. Dealing with all these issues no matter how many sources we combine is an efficient manner to provide the best information in the process of decision making. The next chapter deals with data clustering and analysis based on a dataset provided from Stack Overflow. We propose this analysis in order to identify who uses question answering communities, and the different profiles in order to detect experts.

4
Data Analysis

Prudens quaestio dimidium scientiae.
Half of science is asking the right questions.

Contents

4.1 Introduction

Data mining is defined as the application of data analysis in order to discover non trivial models in large datasets (Fayyad et al., 1996).

Several algorithms and techniques have been proposed to formalize the exploration of new models, build more efficient ones and measure differences between data sets.

The data mining process consists of three main steps: pre-processing of data, discovery of usage patterns and analysis of results. The data pre-processing phase is often the most labor and time-consuming task, due in particular to the lack of structures and the large amount of noise existing in the used data. For the next phases, we found these tasks on data analysis techniques. From this context, these techniques have been developed since the 1950 encouraged by the emergence of computer science. Nowadays, data analysis has gained more and more popularity due to the huge amount of data we have to deal with. (Tukey, 1977) identified two major categories of data analysis techniques: exploratory and confirmatory. On one hand, for the exploratory (or descriptive) analysis, the data analyst does not have an *a priori* knowledge about the model being investigating. On the other hand, for the confirmatory (or inferential) technique, the data analyst wants to support the validity of the model of the manipulated data. Several statistical methods have been proposed to investigate and analyze data. We can enumerated principal component analysis, analysis of variance, linear regression, discriminant analysis, etc.

Clustering algorithms target to regroup objects in homogeneous classes founded on their characteristics as presented in (Cleuziou, 2004). By homogeneous classes we mean regrouping objects that are close to each other while separating them from different objects that are not similar.

This chapter is organized as follows: The second section details the principle component analysis. The next section reviews the mixed clustering which is composed of two major steps the hierarchical clustering methods and partitioning methods. The fourth section presents a description on the data manipulated in this thesis, we also propose a detailed data analysis. Before concluding, the fifth section details the data analysis performed on real data provided by Stack Overflow.

4.2 Principal component analysis

Principal Component Analysis (PCA) was first introduced by (Pearson, 1901). For the PCA, the number of principal components is less than or equal to the number of original variables or the number of observations. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to

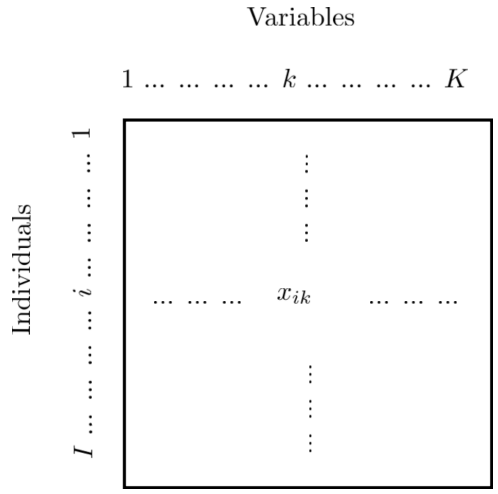


Figure 4.1: Data representation in PCA

the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

The main variants of the PCA are the differences of transformations in the dataset. It can be exploited in several fields of application. There are actually two ways to use it.

- for the study of a given population by seeking to determine the typology of individuals and variables.
- to reduce the dimensions of the data without significant loss of information.

The data for the PCA are usually presented in the form of a table as described in Figure 4.1. PCA must evaluate the similarities between individuals and the links between variables.

Definition 1 *Two individuals are similar, or close, if they have similar values for all variables. This definition implies a notion of proximity which results in a distance. Thus, we define the Euclidean distance between two individuals i and j by:*

$$d^2(i,j) = \sum_{k \subseteq K} (x_{ik} - x_{jk})^2 \tag{4.1}$$

Definition 2 *Two variables are related if they have a high linear correlation coefficient. The linear correlation coefficient is expressed by the following equation:*

$$r(k,h) = \frac{cov(k,h)}{\sqrt{var(k)var(h)}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{ik}-\overline{x_k}}{s_k}\right)\left(\frac{x_{ih}-\overline{x_h}}{s_h}\right) \tag{4.2}$$

where $\overline{x_k}$ and s are respectively the mean and the standard deviation of the variable k.

PCA can be used to reduce the dimensions of a data set. It reduces the data down into it's basic components.

4.3 Mixed Clustering

In this section we recall some of the most used clustering algorithms. The main reason for enumerating many clustering methods is due to the fact that a cluster is not precisely defined (Estivill-Castro and Yang, 2000). Accordingly, several methods have been proposed, each one uses a different induction principle. In (Fraley and Raftery, 1998), the authors proposed to divide clustering techniques into two major groups: hierarchical methods and partitioning methods. When we are in presence of a large number of objects ($> 10^3$), it is impossible to use hierarchical classification methods directly. The best way to get around this issue, is to combine both non hierarchical and hierarchical techniques.

4.3.1 Hierarchical methods

In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters (Maimon and Rokach, 2005) as described in Figure 4.2. Two approaches are used to perform for hierarchical clustering:

- Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster. At each stage of the classification process, the two closest clusters in the sense of an aggregation measure are merged. The process stops when the two remaining clusters merge into the single cluster containing all the individuals.
- Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. The division of a cluster is carried out in a way that the aggregation measure between the two descending clusters is as large as possible, so as to create two well separated clusters.

The result of hierarchical methods is a dendrogram where the nested grouping of objects changes. We obtain a clustering of the data by cutting the dendrogram according to a similarity level.

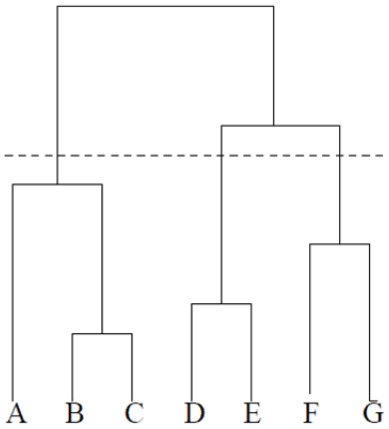


Figure 4.2: Example of a dendogram

Hierarchical clustering methods can be further splitted based on the way how the similarity measure is calculated. In (Jain et al., 1999), the authors proposed three approaches:

- Single-link clustering: also called the correctness, the minimum method of the nearest neighbor method. It considers the distance between two clusters to be equal to the shortest distance from any object of one cluster to any object belonging to the other cluster.
- Complete-link clustering: also called the diameter, the maximum method. It considers the distance between two clusters to be equal to the biggest distance from any object of one cluster to any object of the other cluster (King, 1967).
- Average-link clustering: also called minimum variance method. It considers the distance from any object of one cluster to any object of the other cluster. This kind of algorithms can be found in (Murtagh, 1983).

4.3.2 Partitioning Methods

Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning. Such methods typically require that the number of clusters will be pre-set by the user. To achieve global optimality in partitioned-based clustering, an exhaustive enumeration process of all possible partitions is required. Because this is not feasible, certain greedy heuristics are used in the form of iterative optimization according to (Maimon and Rokach, 2005). The number of classes in the partition to be generated must be fixed at the start.

An optimal partition can be obtained from the exhaustive enumeration of all the partitions, which however becomes prohibitive in terms of computation time. As an alternative solution to this problem, partitioning methods based on the iterative optimization of the criterion make it possible to obtain distinct groups in a reasonable calculation time. These optimization methods use a reassignment in order to iteratively redistribute the individuals in K classes.

K-means

K-means is a very popular algorithm used for data clustering. The K-means requires three user-specified parameters: number of clusters K, cluster initialization, and distance metric. The most critical choice is K. A wrong choice of the value of K may lead to incorrect clustering. Thus, the K-means algorithm is run with different values and the partitions the most relevant partitions are selected.

This algorithm is typically used with the Euclidean metric for computing the distance between points and cluster centers. As a result, K-means finds spherical or ball-shaped clusters in data.

The main steps of the algorithm are described in the following:

1. Select an initial partition with K clusters.
2. Generate a new partition by assigning each pattern to its closest cluster center.
3. Compute new cluster centers.
4. Repeat steps 2 and 3 until the obtained partitions are relevant.

K-medoids

The K-medoids algorithm is a clustering algorithm related to the k-means algorithm and the medoidshift algorithm. Both the k-means and k-medoids algorithms are partitional (breaking the dataset up into groups). K-means attempts to minimize the total squared error, while k-medoids minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the K-means algorithm, K-medoids chooses datapoints as centers (medoids or exemplars). K-medoids is also a partitioning technique of clustering that clusters the data set of n objects into K clusters with K known a priori. A useful tool for determining K is the silhouette.

It could be more robust to noise and outliers as compared to K-means because it minimizes a sum of general pairwise dissimilarities instead of a sum of squared Euclidean distances. The possible choice of the dissimilarity function is very rich but in our applet we used the Euclidean distance.

A medoid of a finite dataset is a data point from this set, whose average dissimilarity to all the data points is minimal *i.e.* it is the most centrally located point in the set.

4.3.3 Methodology of mixed classification

As the hierarchical techniques are very greedy for time computing and for the partitioning techniques we have to determine from the beginning the number of classes. The best compromise is to use a combination of these two methods. The mixed classification is a method that seeks to group the advantages of both hierarchical and partitioning methods and to overcome their disadvantages.
To do so, we need to take into consideration the following steps:

1. Use a partitioning method by choosing the number of clusters.
2. Construction of a dendrogram (also called tree) from the k classes formed at step 1. Cutting of the dendrogram into an optimal number of classes.
3. Consolidation of the partition obtained at step 2.

4.4 Data Analysis

In this section we will perform an analysis on the dataset provided from Stack Overflow. The main goal of this data analysis is to condense the information contained in a large number of original variables into a smaller set of new composite dimensions, with a minimum loss of information. This analysis will allow us to reveal important features present in our large dataset. It will also help us to explore relationships that were initially unsuspected. This analysis is an important step that has to be performed in order to better understand the available data and identify the most significant variables. The remainder of this section is organized as follows: first, we describe and present some statistics about the data that will be used in this study. Next, we will perform a principle component analysis in order to investigate the relationship between the variables and reduce them into a set of 'artificial' variables also called components. After the PCA, we achieve a mixed classification using a clustering based on a K-means and a hierarchical ascendant classification. All these analysis will allow us to distinguish between different categories of user in question answering communities.

4.4.1 Dataset description

In this thesis, we used data provided by Stack Overflow from December 2013 to March 2015[20]. The data set counts over 2 Million users, 2.5 Million answers and 1.7 Million

[20] https://archive.org/download/stackexchange

questions. In Table 4.1, we present some statistics about the dataset provided by Stack Overflow and used for these experiments.

Table 4.1: Statistics

	Mean	Standard deviation	Minimum	Maximum
Reputation	264.98	$3.7189 * 10^3$	1	799760
Views	30.734	877.81	0	1013100
UpVotes	26.304	195.69	0	136650
DownVotes	2.7463	108.896	0	7572500
Number of Answers	0.6647	9.1757	0	5595
Number of Questions	0.3629	2.1626	0	316
Number Accepted Answers	0.13	2.769	0	1781
Time Activity	556.5468	607.6323	1	2571
Age of the user	31.73	8.229	13	95

4.4.2 PCA on the dataset

As we are dealing simultaneously with a large number of quantitative variables. For this data analysis we will only focus on six major features: Number of views, upvotes, downvotes, questions, answers and accepted answers. As the reputation is only a measure of popularity we will not take it into account. For the "age" as users do not always fill the right information about them, we will not treat these information. The main difficulty of PCA arises from the fact that the studied individuals are no longer represented in traditional representation, but in a space of mutli-dimensions.
The study of the correlation allows us to measure the intensity of relation that may exist between variables. It is very useful because it can predict a relationship that can be exploited for the data analysis. The correlation between the attributes of each user are presented in Table 4.2. The closer the coefficient is to the extremes -1 and 1, the greater the linear correlation between the variables. We simply use the term "strongly correlated" to qualify the two variables. Thus, we notice that in Table 4.2, there is a strong correlation between the number of answers and the number of accepted answers with a correlation of 0.950. These two variables are the most correlated ones.

Table 4.3 represents eigenvalues related to our dataset. Each eigenvector is a new basis vector. It depends on all dimensions from the original input space. The corresponding eigenvalue tells how much variance that particular variance explains from the total.

Based on these values, we can estimate the explained variance. The results are described in Figure 4.3. It is the percentage of the total variance explained by each principal component. Here, we can notice that the first component summarizes over 89 % of information described.

Table 4.2: Correlation between variables of the dataset

	Views	UpVotes	DownVotes	Nb Qu	Nb Ans	Nb AccAns
Views	1					
UpVotes	**0.6085**	1				
DownVotes	**0.7020**	0.4538	1			
Nb of Questions	0.0026	0.0116	0.0017	1		
Nb of Answers	0.0090	0.0055	0.0090	0.0615	1	
Nb Acc Answers	0.0072	−0.005	0.0071	0.0411	**0.9560**	1

Table 4.3: Eigenvalues of the variables

	Views	UpVotes	DownVotes	Nb Qu	Nb Ans	Nb AccAns
Views	**0.9071**	−0.4185	−0.0438	0.0015	−0.001	0.0001
UpVotes	0.4097	**0.9021**	−0.1352	−0.0037	0.0002	−0.0002
DownVotes	0.0961	0.1046	**0.9899**	0.0005	0	0.0001
Nb of Questions	0	−0.0001	0	0.0291	**0.9854**	0.1678
Nb of Answers	0.0002	−0.0036	−0.0008	**0.9426**	0.0288	−0.3326
Nb Acc Answers	0.0002	−0.0016	−0.0004	0.3325	−0.1678	**0.9280**

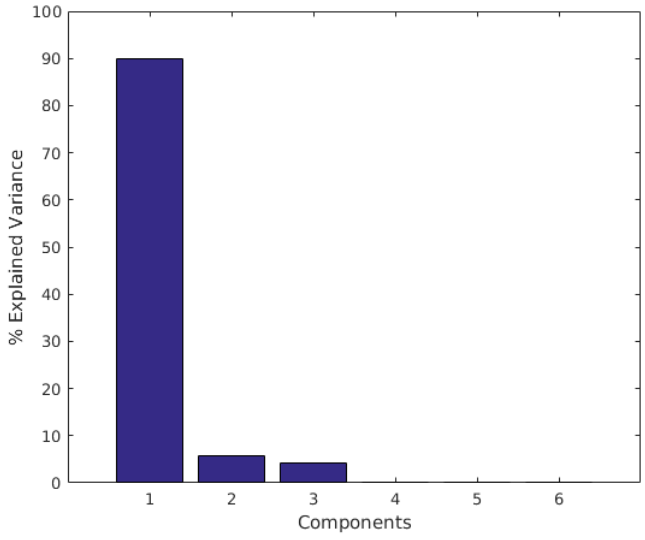


Figure 4.3: Explained Variance of each component

4.4.3 Estimation of the optimal number of Principal Components

Before starting the principal component analysis we have to determine their optimal number. To do so the literature proposes two major methods. The first one is based on the eigenvalues and the second one is founded on the multivariate k correlation index.

The techniques based on eigenvalues are the Average Eigenvalue Criterion (AEC, also known as Kaiser's criterion) and the Corrected Average Eigenvalue Criterion (CAEC). (Henry F, 1960) proposed two simple methods based on eigenvalues are:

- The Average Eigenvalue Criterion (AEC, also known as Kaiser's criterion) defines as significant only the components with eigenvalue greater than the average eigenvalue.
- The Corrected Average Eigenvalue Criterion (CAEC) is the same as AEC, but simply decreases the rejection threshold by multiplying the average eigenvalue by 0.7.

For the techniques using the multivariate k correlation, the author, (Todeschini, 1997) distinguishes between a linear function (KL) and a non-linear one (KP). They are described in the following equations:

$$KL = int\left[1 + (p-1) * (1-k)\right] \tag{4.3}$$

$$KP = int\left[p^{(1-k)}\right] \tag{4.4}$$

where p is the number of original variables and int indicates the nearest integer upper value.

In both equations (4.3) and (4.4) the results equal 1 when $k = 1$ (all the original p variables are mutually correlated, so one component is retained) and equal p when $k = 0$ (all the original variables are orthogonal, so all the components are retained).

While KL gives the maximum number of theoretical significant principal components, under the hypothesis that the information in the data is linearly distributed. KP evaluates the safest minimum number of significant components under the suspicion that the information in the data decreases in a steeply way.

We apply the four techniques presented to our randomly selected 50.000 users in order to obtain the p-optimal number of components. The results are shown in Figure 4.4.

The AEC proposes only 2 components while both of CAEC and KP estimate that 3 is the optimal number of components. The method using KL considers the number

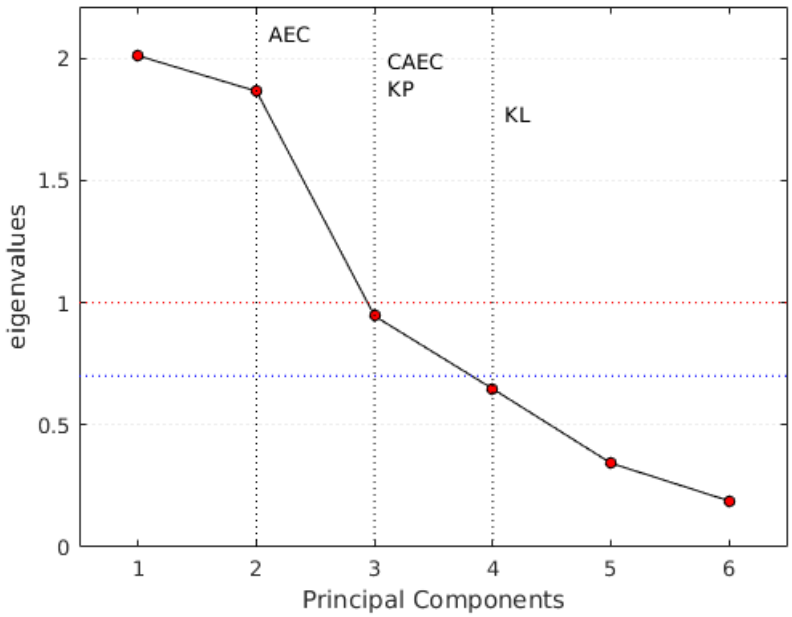


Figure 4.4: Estimation of the optimal number of components

of components as 4. As KL gives the maximum number, this technique may be too optimistic.

The association between the components and the original variables is called the component's eigenvalue.

Based on Figure 4.4 we will assume that we have three principal components for the rest of this thesis.

4.4.4 Results on components

The factors analysis is a way to fit a model to multivariate data to estimate just this sort of interdependence. In a factor analysis model, the measured variables depend on a smaller number of unobserved (latent) factors. Because each factor might affect several variables in common, they are known as common factors. Each variable is assumed to be dependent on a linear combination of the common factors, and the coefficients are known as loadings. Each measured variable also includes a component due to independent random variability, known as specific variance because it is specific to one variable.

Thus, the first step is to determine a new set of orthogonal axes based on the available data. This is achieved by identifying the direction of maximal variance through the coordinates in the 6 dimensional spaces of the dataset. Table 4.4 describes the

Table 4.4: Correlation between variables and components

	PC 1	PC 2	PC 3
Views	**0.5346**	0.3470	−0.0063
UpVotes	**0.5339**	3360	−0.0027
DownVotes	0.3701	0.2427	−0.0203
Nb of Questions	0.1386	−0.2359	**-0.9611**
Nb of Answers	0.3708	**-0.5738**	0.1684
Nb Acc Answers	0.3681	**-0.5682**	0.2180

correlation between, the variables and our three principle components. We have to investigate which variable is greater from zero in either a positive or negative direction. Thus, these variables are correlated with each component.
The first component is correlated with the two first variables: Number of Views, Upvotes and DownVotes. This component can be recognized as a measure of how a user is considered within the community with the number of positive or negative votes gained in the platform respectively 0.5339 and 0.3701 and the number of times the profile has been viewed with a correlation of 0.5346.
The second principal component is strongly and negatively correlated with two variables: the number of answers and the number of accepted answers respectively with −0.5738 and −0.5682. This component can be seen as a measure of the how active a user is and how he is willing to contribute and help the other members of the platform with helpful and high quality answers. This second principal axis also takes into account the number of positive votes and views. Thus, these variables are the reflection of a user's popularity and reward based on the quality of his posts.

Finally for the third principle component, we notice that this axis is only correlated on a negative direction to the number of questions posted in the platform with a value of −0.9611.

The main purpose of the PCA is to reduce the information contained in the dataset based on an analysis of the correlation between variables and graphical visualization of the distances between the individuals. The circle of correlations is the projection of the variables on the principle components. Actually, the closer the variables are to the circle, the better they are represented. The more the angle between two variables is small, the more they are correlated. On one hand, when the variables are opposed graphically on the plot, the angle between them tends to -1 and thus the more the two variables are negatively correlated. On the other hand, when the angle between the variables is close to 0, there is no linear correlation between the two variables.

The next step of this analysis is the projection of the variables according to the three principal components. Figures 4.5, 4.6 and 4.7 respectively describe the illustration of the circles of correlations of the variables presented earlier in Table 4.4.

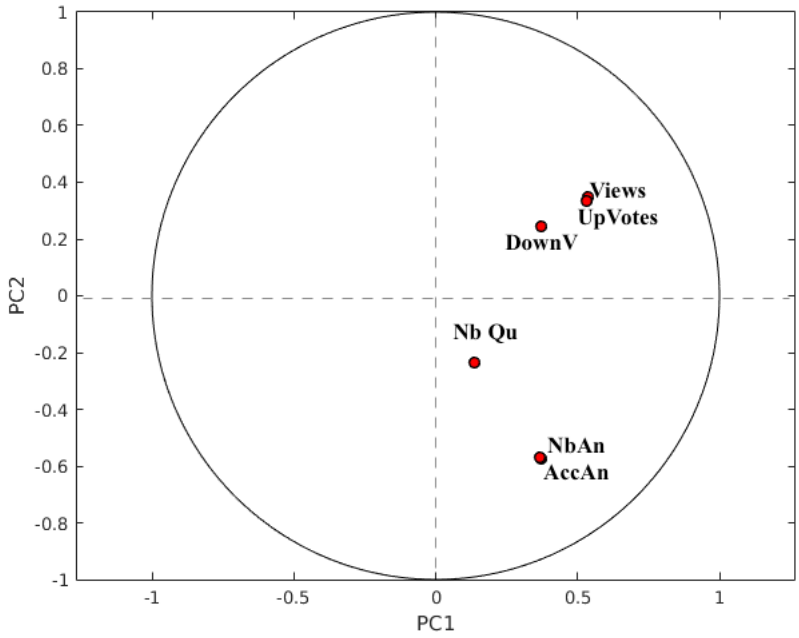


Figure 4.5: Projection of variables on components 1 and 2

Figure 4.5 shows the plot of all the six variables illustrating how each one of them contributes to the construction of the two first components.

All variables are represented in this bi-plot by a vector, and the direction and position of the plots indicate how each variable contributes to the two principal components in the plot. For example, the first principal component, on the horizontal axis, has positive coefficients for Views, UpVotes.

We can see that in figures 4.5, 4.6 and 4.7 the number of UpVotes and views are very correlated and important for the constitution of the first components. The number of Answers and Accepted Answers are also very close besides to their contribution to the second principle components. For the number of questions, we can see that this variable is very close to the circle of correlation in figure 4.6, making it very important for the definition of the third component.

4.4.5 Results on individuals

We present the projection of the individuals on the three principle components presented in the section below.

As shown in Figure 4.8, the results of the projection of the users in the first and second principle components show that the objects are centered in the original points

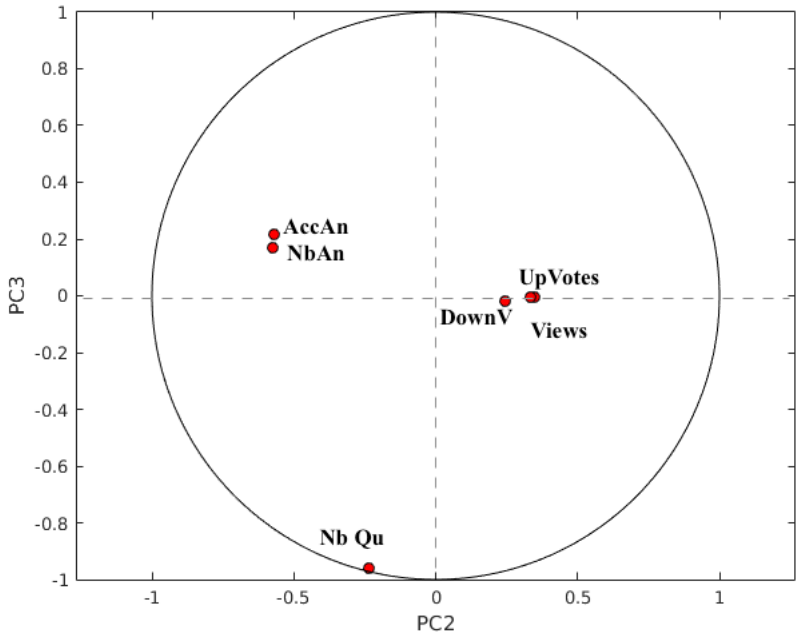


Figure 4.6: Projection of variables on components 2 and 3

of the axis. However we observe that they follow in the positive direction the variables Views and upvotes. On the negative direction they are oriented according the number of posted answers and accepted answers.

We can see from Figure 4.9 that a lot of the users projected on axis 2 and 3 are concentrated at the center of the plot. However, we can find some outliers that are projected in the negative direction of axis 2. The users present in these axes seem to be providing an important number of answers which some of them are accepted and chosen as the best answers.

In Figure 4.10 we can observe that although users are centered, several users are dispatched a long the first axis and in both sides of the third axis (positive and negative directions). These users seem to be posting a lot of questions in the platform and receive positive votes for that.

4.4.6 Hierarchical clustering of Data

In this section we proceed to a clustering using an Agglomerative Hierarchical Clustering as described in section 4.3.1. However, due to the large size of the manipulated dataset and the complexity of this technique, we decide to perform a mixed classification. A mixed classification is divided on two major steps:

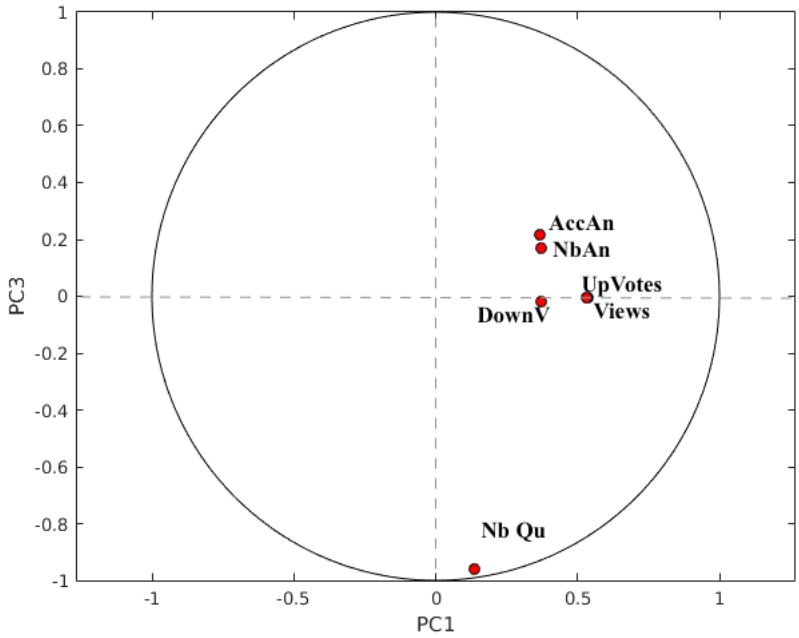


Figure 4.7: Projection of variables on components 1 and 3

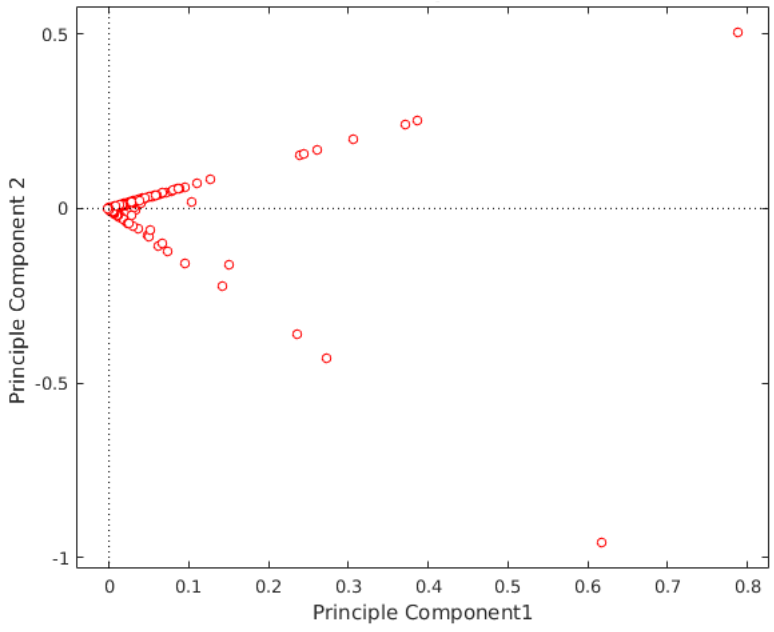


Figure 4.8: Projection of individuals on components 1 and 2

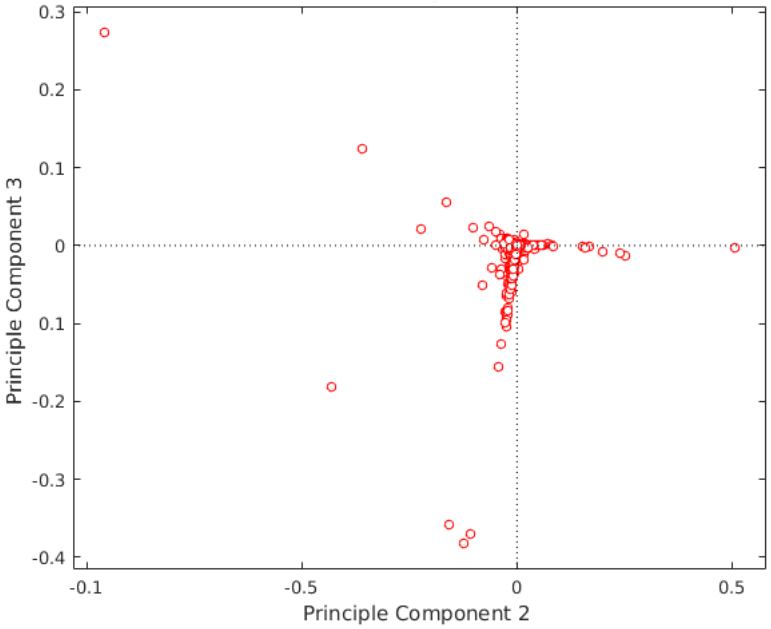


Figure 4.9: Projection of individuals on components 2 and 3

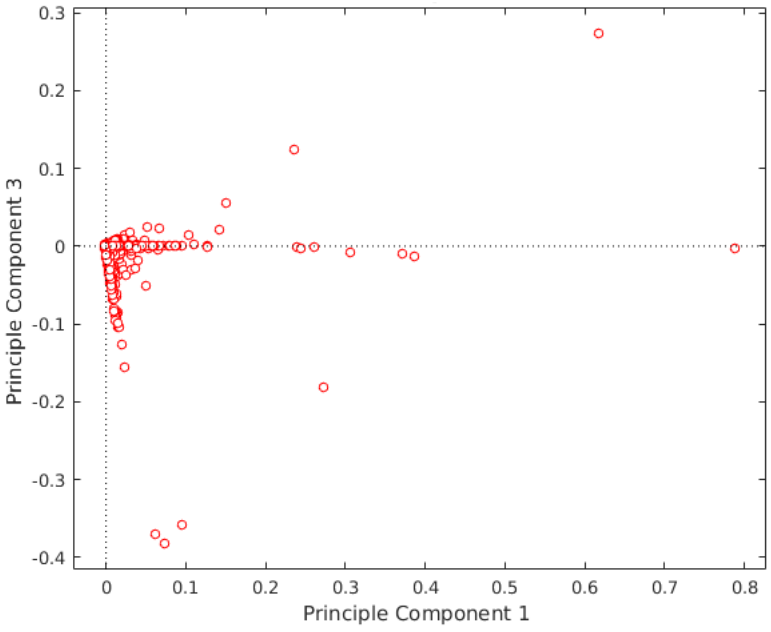


Figure 4.10: Projection of individuals on components 1 and 3

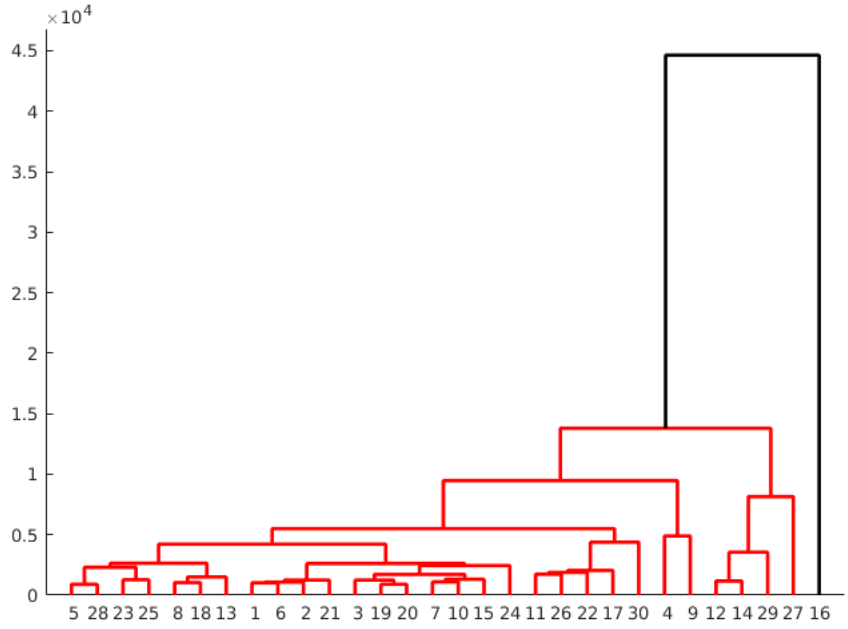


Figure 4.11: Dendogram of hierarchical clustering

- A clustering technique such as K-Means, K-Medoids
- An Agglomerative Hierarchical Clustering

We perform a K-Means clustering on a 10.000 randomly selected users with $K = 1000$. We obtain a clustering with a thousand classes. For each of them we calculate the center of the cluster. Next we use these results and perform a hierarchical clustering

First we start with 1000 clusters, then we merge them based on an euclidean distance.

As the reputation is a direct indicator of a user's popularity in the community we did not include it to investigate the expertise. In these experiments we only use the following attributes: Number of positive and negative votes, number of questions, answers, accepted answers and time activity. We do not take into consideration the reputation.

In figures 4.12, 4.11 and 4.13, we present the dendograms obtained by hierarchical clusterings on the 1000 clusters given by the k-means, for the six variables used in the PCA, questions and answers and finally for reputation. We zoom the displays to the last 30 clusters.

The cluster defined by the number 16 is considered as an unique cluster composed

by only one user. This user has a high number of views and positive votes. Clusters $27, 29, 14$ and 12 do also have a great number of profile views and upvotes. Clusters 9 and 4 have the same characteristics as the previous clusters except that they have a lot of negative votes. These individuals even though they are very popular they are not very active in the platform (number of posted questions and answers are relatively small around 100). Branches $15, 10$ and 7 are characterized by a high number of views, an average number of positive votes and a high number of negative votes (around 100). The majority of their contributions is answers posting. It seems that their posts have a low quality which is justified by a such number of downvotes. We witness the same phenomenon for the cluster number 24.
For the clusters $21, 2, 6$ and 1, users do not have a very high number of views or upvotes, yet when we focus on the number of posts we notice that some of them are very active in the platform they have a great number of accepted answers. However, a large part of the population is not very active within the community. Thus, the active contributors are embedded in the huge number of users.
We notice a controversy issue, when we deeply analyze the results of Figure 4.11. Among all the attributes describing users in Stack Overflow, the number of views has more impact that the others. This directly impacts the results of our analysis. Moreover, we discovered some inconsistencies in the data provided by the platform. Some users have a high number of accepted answers, although if they have a null value of positive votes or a very small reputation number. This can not be possible considering the gratification system of Stack Overflow presented in Table 3.2. In case of an answer chosen as the best, the answerers gains at least 15 points of reputation and the asker 2 points.
Taking into account the fact that the number of profile views is closely related to the popularity of a user, we decide to perform a second hierarchical classification focusing only the number of posts: questions, answers and accepted answers.

The obtained results are displayed in figure 4.12.

We can divide the results of the hierarchical classification into three or four clusters based on where we cut the dendrogram.

The behavior of users composing the first cluster is different compared to the two other clusters. They are numerous. They participate moderately within the community. Most of their contributions are questions asked from time to time. Their questions seem to be considered as helpful and are often positively rated.

Users belonging to cluster 2 are active users, they contribute a lot compared to the first cluster. They post a lot of questions and answers. However, they do not have a high ratio of accepted answers per total number of answers. When we examine deeply their scores, we see that most of them have more negative votes than positive. Even though most of them are not very reliable they may have some knowledge when we take a closer look at the quality of their posts. We can also notice that users of cluster 2 seem to post very often keeping the community active. Even though the cut of the

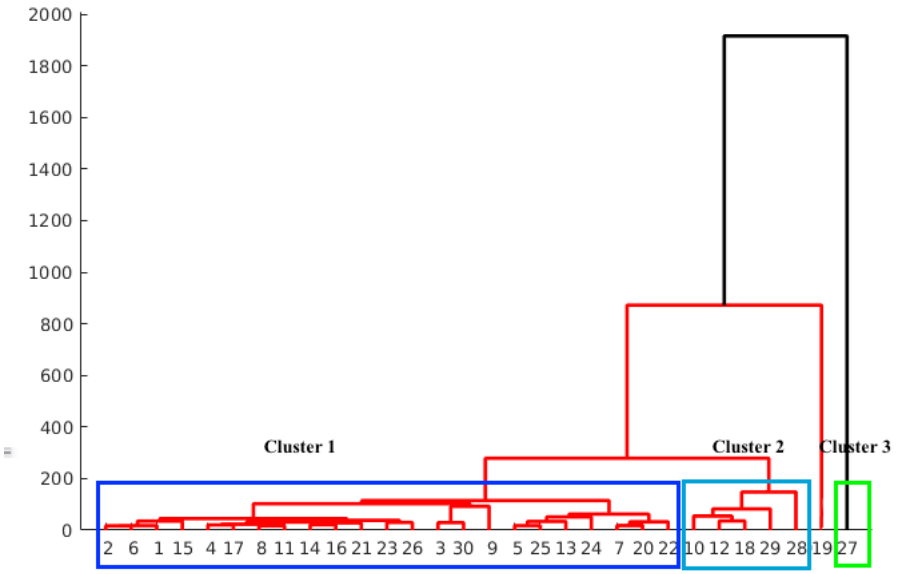


Figure 4.12: Dendogram of hierarchical clustering of Questions and Answers

resulting dendogram defines 4 clusters, we decide to not take into account the node 19. As this node is composed by only one user.
This decision is based on the behavior of this particular user which is examined in details. This person is considered very active too but not as much as node 27, with 333 best answers. We also notice that this user seems to be asking more questions than the other persons in this cluster. Inside this group of users we can witness that there are two main trends: skilled and unskilled answers. Thus, considering that the number of their contributions is quite similar we include them to the same cluster. The active and skilled users are very close to the third cluster, we take for example node number 19.

Next we focus on analyzing the third cluster's results. This node is composed by two users, we can see that they are highly active with an average number of answers 1421 and accepted answers (494). Yet, these individuals do not ask much questions making them a valuable user especially for providing answers to their peers in the community.

Thus, based on the results provided by Figure 4.12, we can deduce that there are three major categories of users in question answering communities.

Figure 4.13 represents a hierarchical clustering on the reputation only. The results obtained show that we have two major trends according to the value of the user's reputation. Cluster 1 contains users whom reputation is between 1 and 125.000, while cluster 2 contains user with a value greater than the first cluster. Even though we are dealing with a high number of reputation, the present number of questions and

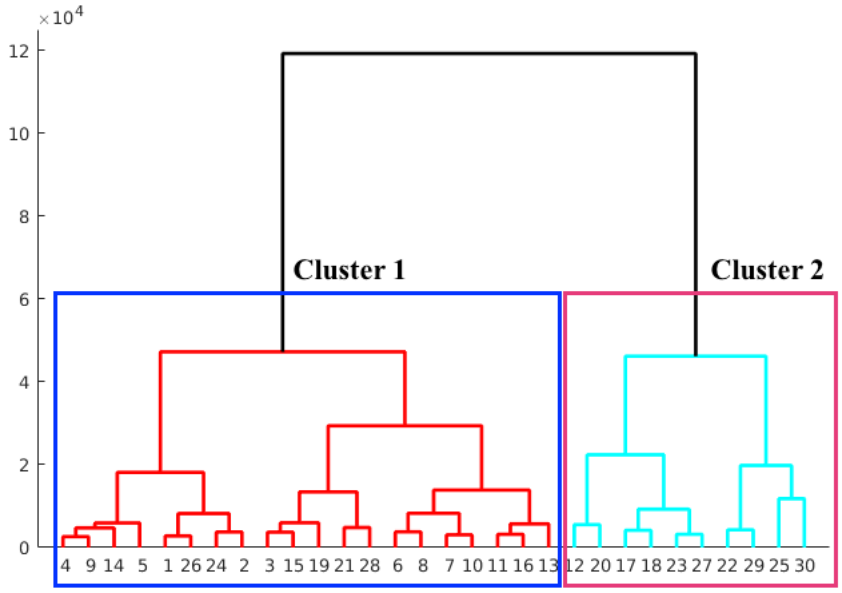


Figure 4.13: Dendogram of hierarchical clustering of the reputation

answers posted in the platform are inferior. Consequently, due to this fact, we do not consider reputation for clustering users and especially according to their popularity in the community.

4.4.7 Users Clustering in Stack Overflow

With the increasing popularity of Q&A C, we can see that users behave differently. Some of them are present in the platform only to obtain information when needed. Some other users are just interested by gaining popularity and being recognized in their community by their peers. These reputation collectors are not always good contributors. Their only objective is to post as much as possible in hope of gaining popularity.

However, we can also detect another category of users. They are involved inside the community and try to help other members to solve their issues by providing answers. Their answers are usually well expressed, precise and helpful.

Based on the analysis provided in the section below, we can distinguish between three trends of users based only on their activities. The number of posted questions and answers is very important, it reflects how involved a user is in the community. The number of accepted answers in an indicator of how helpful and precise a contributor can be while responding to a question. The number of positive votes reveals how the answer is perceived by the members of the community. The high number of downvotes implies the low quality of posts.

Considering both the principle components analysis and the hierarchical classification we can identify three groups in order to classify users in Stack Overflow:

- **Occasionals (O)**: these users represent the major part of members on the platform. They do not have a lot of knowledge. They occur occasionally only when they need an answer to a specific question that have not been treated before. They do not have a lot of positive votes. These users are characterized by a very scarce activity (few questions and answers posted). These users are the most numerous in the community. They represent the Cluster 1 in Figure 4.12.

- **Apprentices (A)**: these users may have some expertise in a given topic. They aim to increase their reputation. To do so, they post a lot of answers that are not always very useful. The quality of their posts is not guaranteed and their answers can be down-voted. These users are quite active in the platform. They aim to gain knowledge and ability to post helpful answers. Their main purpose in Stack Overflow is to have as much reputation points as possible making them to have more notoriety among their peers. These users represent the Cluster 2 in Figure 4.12.

- **Experts (E)**: these users are very reliable and recognized by the community. They provide an important number of useful answers that are chosen as the best ones. They are very active in the platform and guarantee a high quality content. These experts are very scarce and they represent the Cluster 3 in Figure 4.12.

4.5 Conclusion

In this chapter, we mainly focus on the problem of of data analysis. We introduced several clustering techniques reported in the literature. We also presented the data used for this thesis provided by Stack Overflow. After that, we performed a Principal Component Analysis (PCA) to identify new patterns in the data. We highlighted their similarities and differences.

PCA is well-know to study the interrelations among a set of variables characterizing users in Q&A C. This analysis was made in order to identify the underlying structure of users' attributes. Later, based on this analysis we performed a mixed classification. We identified three main types of users: Occasionnals, Apprentices and Experts.

The next chapter will focus on defining a general measure of expertise based on the theory of belief functions. This measure will allow us to detect experts and classify users according to the clusters presented above using some attributes describing them.

5
User's classification based on a Belief Measure of Expertise

"Understanding our world requires conceptualizing the similarities and differences between entities that compose it."
Tyron and Bailey 1970

Contents

5.1 Introduction

In chapter 3, we presented how uncertainty is modeled using the theory of belief functions and represented by mass functions. In chapter 4, we described some techniques for unsupervised machine learning and the three main classes of users that can be discovered in Stack Overflow. We distinguished between Experts, Apprentices and Occasionnals.

While most of the researches in question answering communities focused on experts detection, some of them considered several types of users and proposed to classify them. (Furtado et al., 2013) discovered 10 classes of users while, (Ma et al., 2015) distinguished between askers, answerers and voters.

This chapter is focused on classifying users in three main classes: Experts, Apprentices and Occasionnals as presented in the last section of Chapter 4. To do so, we propose to measure the general expertise of users based on their activity in Stack Overflow.

The main contributions presented in this chapter are the following: first, we propose a measure of expertise based on the theory of belief functions. This measure combines several information characterizing users in Stack Overflow. We note that the proposed measure can be adapted for other question answering communities. This measure is a global estimation of expertise and is the subject of the paper (Attiaoui et al., 2017a). This chapter is organized as follows: Section 5.2 details the hypothesis that will allow us to distinguish between the three clusters of users. Next, section 5.3 details the statistical model proposed to measure the general expertise of users based on the theory of belief functions. Later in section 5.4, we present the classification of users according to both the credal and pignistic level for decision making.

5.2 Hypothesis for users' modeling in Stack Overflow

In this section, we present the hypothesis that will be used in order to build our model. First we will start by presenting the attributes representing each user in Stack Overflow.

5.2.1 Users' attributes

Among the attributes that can describe a user in Stack Overflow, we select five different and important features characterizing users in this platform:

- **Number of Up Votes** (UV_i): the sum of positive votes collected by posted questions and answers.
- **Number of Down Votes** (DV_i): the sum of negative votes collected by posted questions and answers.

- **Time activity:** time of activity of users from their registration to their last connection.

- **Number of posted questions** ($NbQu_i$): number of questions posted in the dataset during the time activity of a user.

- **Number of posted answers** ($NbAn_i$): number of answers provided in the dataset during the time activity of a user.

- **Number of accepted answers** ($NbAccAn_i$): number the answers chosen as the best answers.

5.2.2 Hypothesis for modeling users in Stack Overflow

According to the previous presentation of the classes of users, we can define the following hypothesis:

Hypothesis 1 *If a user has a high score of positive votes this might mean that this person is an expert rewarded for the questions and answers posted in the platform.*

Hypothesis 2 *If a user has a high score of negative votes this might mean that this person is an apprentice seeking information, and rewarded for posting well asked questions and interesting answers posted in the platform.*

Hypothesis 3 *If a user has a high number of answers posted this can be justified by two facts. First, this person is an expert, providing high quality content. Second, it can be an apprentice trying to become an expert by proving to the community that he/she can be as reliable as an expert.*

Hypothesis 4 *If a user has a high number of questions posted this can represent either an expert or an apprentice. Both of them ask a lot of questions.*

Hypothesis 5 *If a user has a high number of accepted answers this can only represent experts. Experts are frequently chosen as the most helpful answers providers.*

5.3 Belief users' modeling in Stack Overflow

In this section we detail the mathematical model that define the hypothesis presented above.

5.3.1 Definition of mass functions

For each hypothesis, we detail how we build the mass functions in order to represent the data relative to each user. Considering the three classes of users Occasional (O), Apprentice (A) and Expert (E). The frame of discernment is $\Omega = \{O, A, E\}$.
The power set is defined as: $2^{\Omega} = \{\emptyset, O, A, O \cup A, E, O \cup E, A \cup E, \Omega\}$

Each user u is characterized by the following features:

- According to the hypothesis 1, a high score of positive votes (UV) is represented by a mass function on the focal element "**Expert**" (E) and the remainder is given to the ignorance, for a user i:

$$\begin{aligned} m_1^i(E) &= (1 - \alpha_1 \, e^{-\gamma_1 UV_i}) \\ m_1^i(\Omega) &= \alpha_1 \, e^{-\gamma_1 UV_i} \end{aligned} \tag{5.1}$$

- According to the hypothesis 2, a high score of negative votes (DV) is represented by a mass function on the focal element "**Apprentice**" (A) and the remainder is given to the ignorance, for a user i:

$$\begin{aligned} m_2^i(A) &= (1 - \alpha_2 \, e^{-\gamma_2 DV_i}) \\ m_2^i(\Omega) &= \alpha_2 \, e^{-\gamma_1 DV_i} \end{aligned} \tag{5.2}$$

- According to the hypothesis 3, a high number of posted questions is represented by a mass on the union of two classes "**Apprentice** $\cup$ **Expert**". Otherwise, when this value is low it is affected to the "**Occasional**" (O) and the remainder to the ignorance. When a mass is on the union, this means that we can not decide which one of these classes is concerned by the mass. For a user i:

$$\begin{aligned} m_3^i(A \cup E) &= \alpha_3 \left(1 - e^{-\gamma_3 NbQu_i}\right) \\ m_3^i(O) &= \alpha_3 \, e^{-\gamma_3 NbQu_i} \\ m_3^i(\Omega) &= 1 - \alpha_3 \end{aligned} \tag{5.3}$$

- According to the hypothesis 4, a high number of answers is represented by a mass on the union of "**Apprentice** $\cup$ **Expert**" while on the opposite situation the

mass is transferred to the "**Occasional**" and the remainder to the ignorance. For a user i:

$$\begin{aligned} m_4^i(A \cup E) &= \alpha_4(1 - e^{-\gamma_4 NbAn_i}) \\ m_4^i(O) &= \alpha_4\, e^{-\gamma_4 NbAn_i} \\ m_4^i(\Omega) &= 1 - \alpha_4 \end{aligned} \tag{5.4}$$

- According to the hypothesis 5, a high number of accepted answers is represented by a mass on the focal element "**Expert**" and the remainder to the ignorance, for a user i:

$$\begin{aligned} m_5^i(E) &= (1 - \alpha_5\, e^{-\gamma_5 NbAccAn_i}) \\ m_5^i(\Omega) &= \alpha_5\, e^{-\gamma_5 NbAccAn_i} \end{aligned} \tag{5.5}$$

In the previous equations, we fix $\alpha_1, \alpha_5 = 0.9$, $\alpha_2 = 1$, $\alpha_3 = 0.8$ and $\alpha_4 = 0.5$. The values are fixed after several experiments in order to have the best representation of each class of users. These values are used to represent the ingnorance in every mass function as described in (Denoeux, 1995). As the apprentices are modeled only one time as focal element in equation (5.3) unlike experts and occasionnals, we choose to affect the value of 1 to α_2. For γ after several experimentations, we decide to keep it as the maximum value of any attribute divided by 100.

5.3.2 Data aggregation and decision making

Figure 5.1 describes the process of evaluation of the Belief Measure of Expertise.

We combine these mass functions using the Demspter's combination rule presented in equation (3.13). Next we apply the pignistic probability and classify the user into Expert, Apprentice or Occasional.
In order to have a better view of a user's expertise, we have to take into consideration for how long this person have registered to the platform. To do so, we decide to apply the discounting operator (noted α^T) presented in equation 3.17. The temporal discounting applied during the combination process allows us to obtain a measure of expertise that includes all the users' attributes besides a the time spent in the community. The combination process allows us to estimate the actual belief expertise for each user during a period of time. It is expressed is expressed by the following equation:

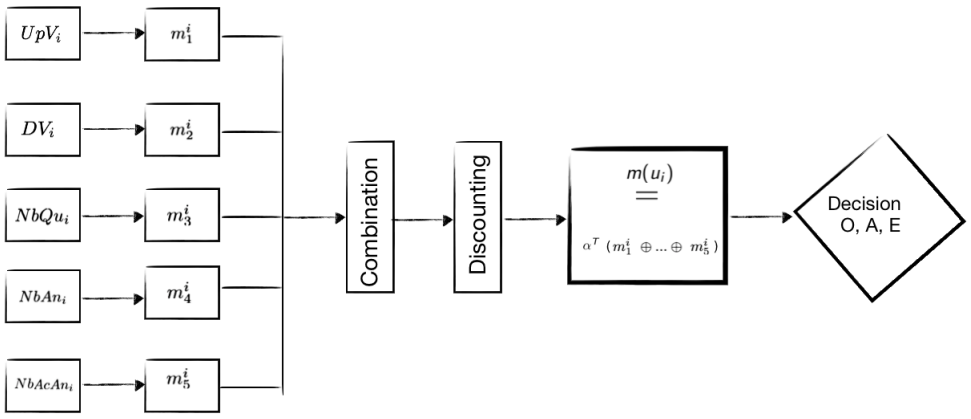


Figure 5.1: Chartflow BME

$$m(u_i) = {}^{\alpha^T}\left(m_1^i \oplus m_2^i \oplus ... \oplus m_5^i\right) \tag{5.6}$$

where α^T is the discounting operator related to the time activity of a user in the platform. The value $\alpha_i^T = 1 - \frac{1}{nd_i}$, where nd_i is the number of days since the user u_i first logged to the platform. The symbol $\oplus$ represents the operator of Dempster's combination.

Once the combination achieved, we perform the pignistic transformation presented in equation 3.18 in order to determine the probabilities of a user being an Occasional, Apprentice or Expert. Thus, the Belief Measure of Expertise (BME) is obtained by the pignistic probability on Experts. It is described by the following equation:

$$BME(u_i) = BetP(E) \tag{5.7}$$

5.4 Users' classification and experts detection

In this section, experiments on real data sets will be performed to show the effectiveness of the proposed measure of expertise. Results will be compared to the Reputation system of Stack Overflow like presented in (Movshovitz-Attias et al., 2013) and a Gaussian Mixture Model (GMM). In (Pal et al., 2012b), authors used GMM as a clustering algorithm to identify clusters among expert users of question answering communities.

They preferred a GMM based method because it overcomes traditional clustering methods such as K-means. In the literature few works proposed a classification of users in question answering communities. Most of the research focused only on the detection of experts. Based on the Belief Measure of Expertise we can propose the classification of users based on three clusters defined in the previous chapter.

5.4.1 Experts detection based on the BME

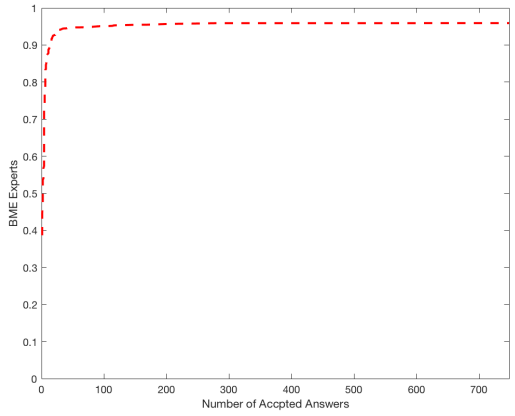
In this section, we show the results of experts detection based on the Belief Measure of Expertise. The number of accepted answers is a very important index on estimating the expertise of a user.

For every user, the BME takes a value in $[0, 1]$. It is the mass allocated to the focal element "Expert". When this value is close to 0 this means that the degree of expertise is weak. Otherwise, when it is near 1 we have a strong belief that this person is an expert. The BME is measured after the combination of the mass functions build from users' attributes and then reinforced by α^T related to the time of activity. After, we classify users according to their pignitic probability described in equation (3.18).

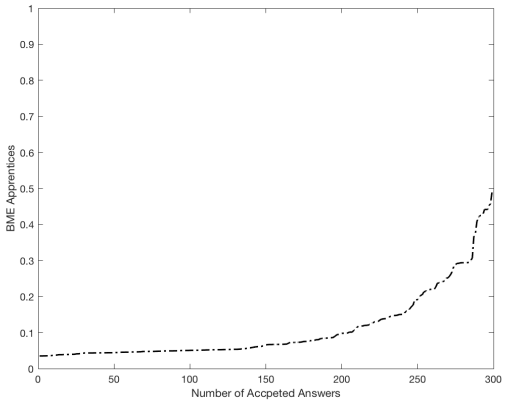
Figures 5.2.a 5.2.b and 5.2.c, show the evolution of the BME according to the number of accepted answers. This feature is considered as one of the most important index about a user's expertise. In figure 5.2.a, we witness that as the number of accepted answers of a user is high, their degree of expertise is increasing and reaching a value greater than 0.9. For the apprentices, the maximal value of their BME when considering the number of best answers given is low with a value of 0.8. The occasionals have the smallest BME resulting from their lack of knowledge and inactivity in the platform. Their former BME is in the interval $[0, 0.3]$.

Therefore, Figures of 5.2.a 5.2.b and 5.2.c reflect the value of the BME according to a very important feature. As the number of accepted answers is increasing and especially for experts, the BME is high and can reach the value close to 1. However, for Apprentices and Occasionals, as they don't have a lot of accepted answers, the value of their BME can be justified by the quality and the number of the questions they posted on the platform. Thus, their BME is small compared to the experts, especially for the occasionals as presented in Figure 5.2.

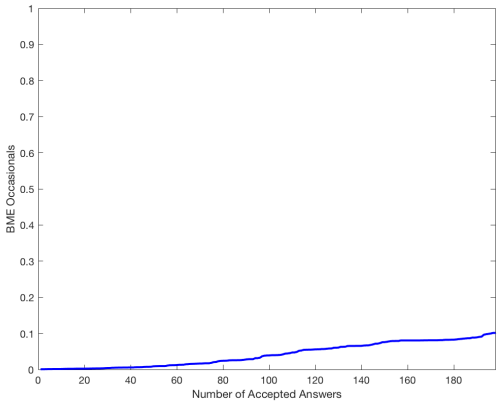
Figures 5.3.a 5.3.b and 5.3.c describe the evolution of the belief measure of expertise according to the number of questions asked in the platform. We notice that experts are the persons that post the highest number of questions reaching even 80 questions for some users during their time of activity in Stack Overflow. We witness the same phenomenon for the number of questions and the number of accepted answers. Nevertheless, when we focus on Figure 5.3.c related to the Occasionals, we notice that



a. BME Experts



b. BME Apprentices

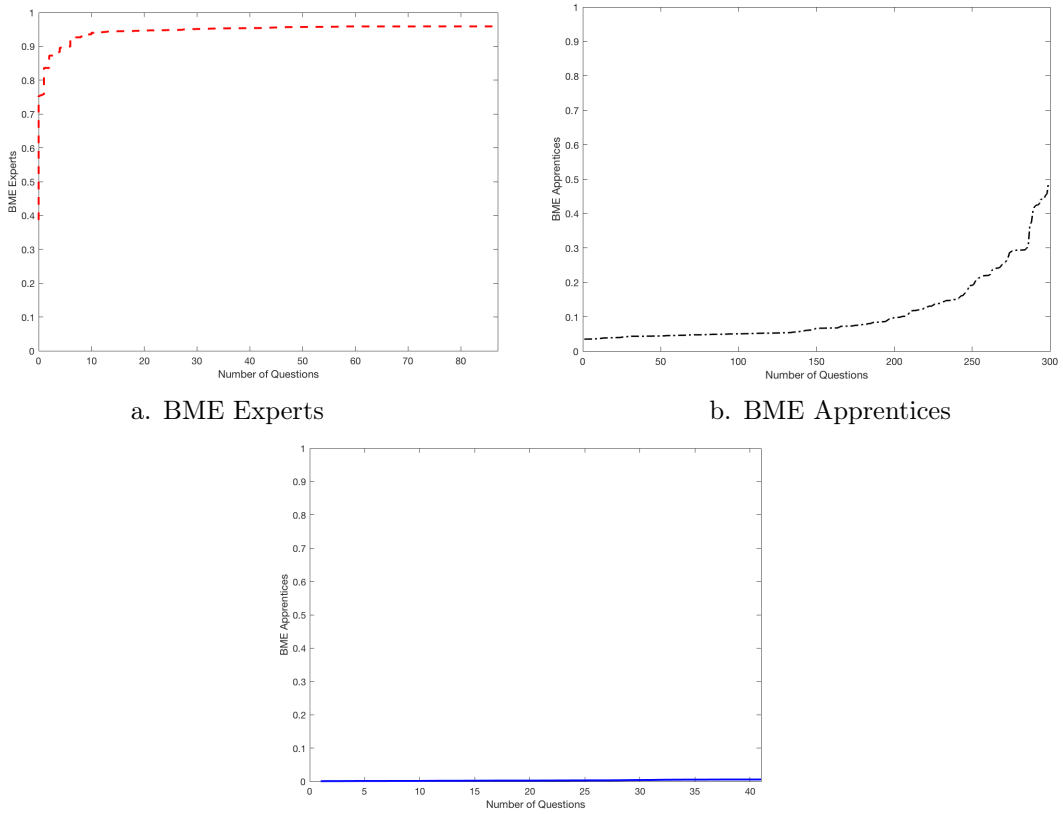


c. BME Occasionnals

Figure 5.2: The BME according to the number of Accepted Answers

users of this class ask punctual questions only when they are seeking information. This leads us to analyze the quality of the questions asked by Occasionals.
Figure 5.4 represents the score gained by Occasionals as an evaluation of the quality of their questions. We notice that some of them seem to be asking interesting questions rewarded by positive votes by the other members of the community despite the fact they are not very active in the platform.

Figures 5.5.a 5.5.b and 5.5.c describe the evolution of the belief measure of expertise according to the score earned by users when providing answers in the platform. We notice that experts as usual have the highest scores especially for their answers. They provide useful and high quality content. The other users of the platform reward them by upvoting their posts allowing them to gain a lot of positive votes as shown in Figure 5.5.a. the more upvotes they obtain, the higher is their expertise level



a. BME Experts

b. BME Apprentices

c. BME Occasionnals

Figure 5.3: The BME according to the number of Questions

Figure 5.6 presents the box plot of the BME for every class. Here, we can see that the degree of expertise of "experts" is the highest. The median value is around 0.7 where some users reach a maximum value of close to 1. However, some users can be considered as experts even though their BME is lower (between 0.37 and 0.5). These users are experts but not as confirmed as the others due to their small time of activity. For the class of apprentices, their BME is smaller than the experts where the median is around 0.15. For the last class of occasionals, their BME is very close to zero. Though, some outliers occur in the apprentices' class with a BME reaching a maximum value of 0.5.

5.5 Evaluation of the clustering

This section concerns the evaluation of the classification methods. This step is essential to estimate and quantify the performances of the algorithms. Several evaluation methods have been presented in the literature to evaluate the quality of the clustering. We

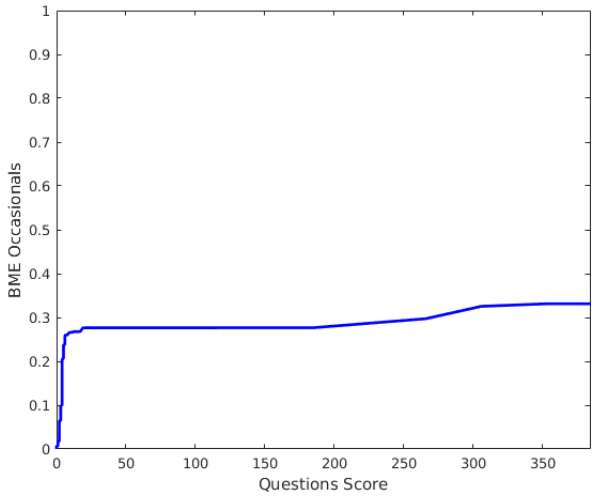


Figure 5.4: Question score gained by Occasionals

chose to use some internal criteria presented in Appendix A. This step will allow us to estimated how well the users are classified based on our belief model and compare the results with the other methods.

We compare the results of our classifier to the reputation system of Stack Overflow and the Gaussian Mixture Model. For the reputation, we use the method described in (Movshovitz-Attias et al., 2013), where experts have a reputation greater than 2400. We fix the experiments of the Gaussian Mixture Model to 3 clusters. The Gaussian Mixture model

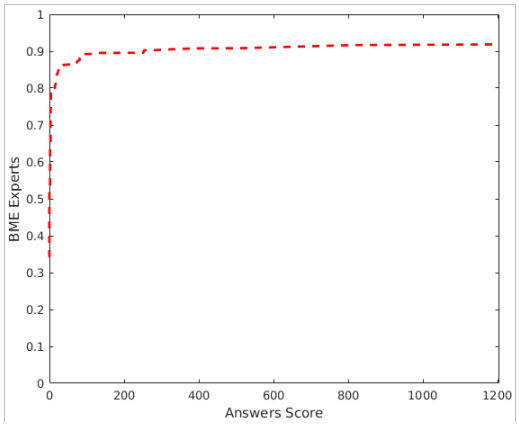
- is considered as the probability distribution that consists of multiple probability distributions
- considers the data as the results of linear combination of several generative Gaussian components

The GMM is expressed by the following equation:

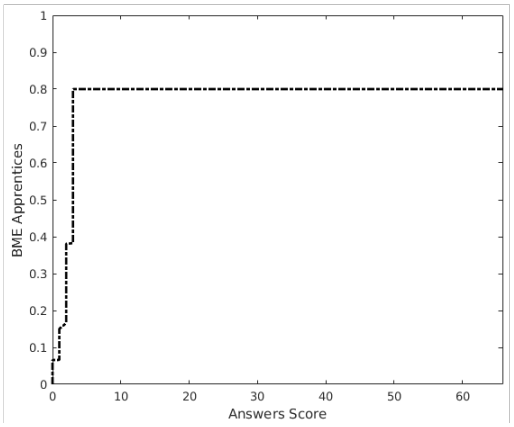
$$f(x) = \sum_{i=1}^{k} w \quad N(x; \mu_i; \Sigma_i) \tag{5.8}$$

where w is the weight, k: number of components, N: the pdf of a Gaussian distribution and μ_i, Σ_i: mean and co-variance of the distribution.

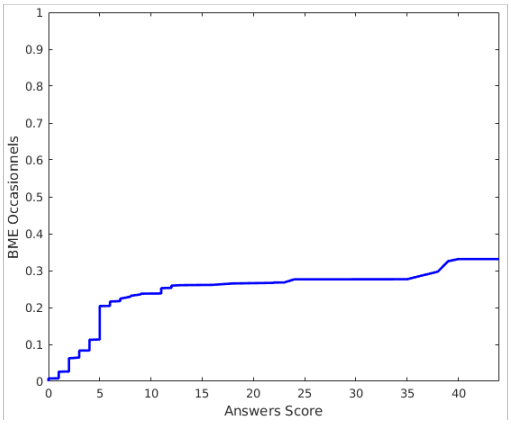
The results of the clustering evaluation are presented in Table 5.1.



a. BME Experts



b. BME Apprentices



c. BME Occasionnals

Figure 5.5: The BME according to the score of answers

The silhouette allows to measure if every object has been well classified or not. The results of the mean silhouette are close to each other around 0.9 for the Reputation and GMM methods. Thus, the latter has the closest measure to 1 (the optimal partition) with a value of 0.98. However, our method presents a value of 0.5563.

The results of the criterion of Davis Bouldin (DB) prove that the proposed approach (BME) presents the most homogeneous classification. The DB focuses on the homogeneity of every class as the best partition has to be the smallest. For the CH, the belief measure of experts (BME) seems to present the best partition. This criterion is closely related to the intra-class inertia. Thus, our method yields better performance than the reputation-based approach. Moreover, the BME presents a better partition when considering the intra-class variance for the Dunn criterion. The BME has better results than the reputation system and GMM. The Dunn criterion examines the distances between the clusters. The value of the reputation based method and the GMM

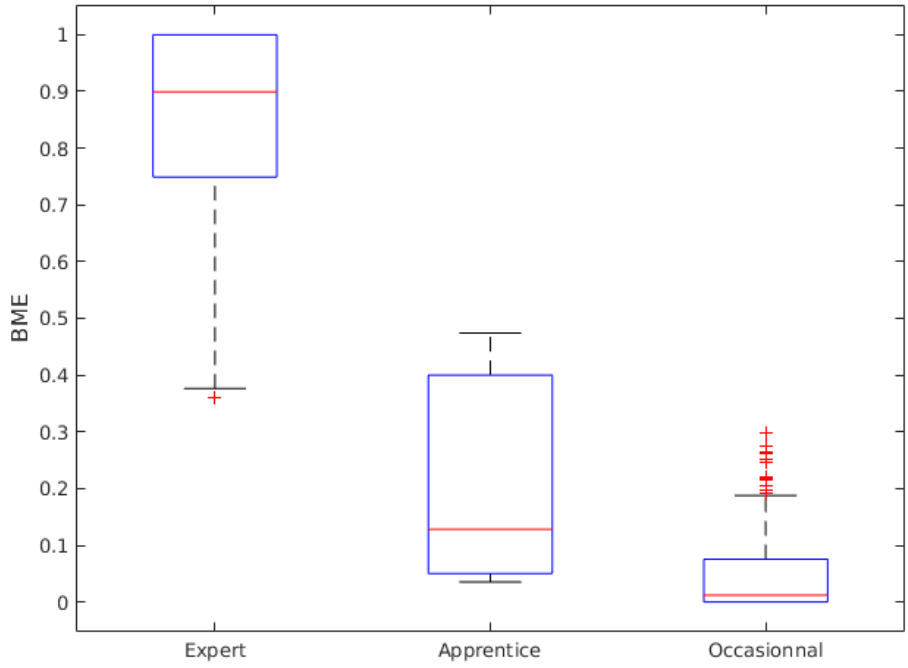


Figure 5.6: Box plots BME Experts, Apprentices and Occasionals

both have small values (< 1) unlike our belief measure of expertise with a value of 5.92. The index of Dunn maximizes the inter-class distance while minimizing the intra-class distance presented in Table 5.1. This index has to be maximized.
For the Root Square (RS) which reflects the degree of difference between classes, the BME has a value of 0.906. As the best partition must be close to 1, the belief expertise measure presents a better partition between the three cluster. This is confirmed with the RMS as its value has to be minimized. Last is RS Error index, where our approach presents the smallest rate of error comparing to the reputation and GMM.

When we compare the results of the several indices of clustering evaluation, the method based on the theory of belief functions outperforms the other approaches. The three clusters generated by our belief model present a more stable and homogeneous grouping of users and especially a better detection of experts.

5.6 Evaluation of the cluster's error

Error bars may show confidence intervals, standard errors, and standard deviations. They are used to show how the data are spread. For this evaluation we will use inferential

Table 5.1: Indices of classification

	BME	Reputation	GMM	Best partition
Mean Silhouette	0.5563	0.9176	0.98	max
DB	0.2216	0.857	1.023	min
CH	40	$6.3 * 10^9$	$4.1 * 10^7$	max
Dunn	5.92	0.139	0.06	max
RS	0.906	0.187	0.123	max
Error RS	0.119	0.317	0.418	min

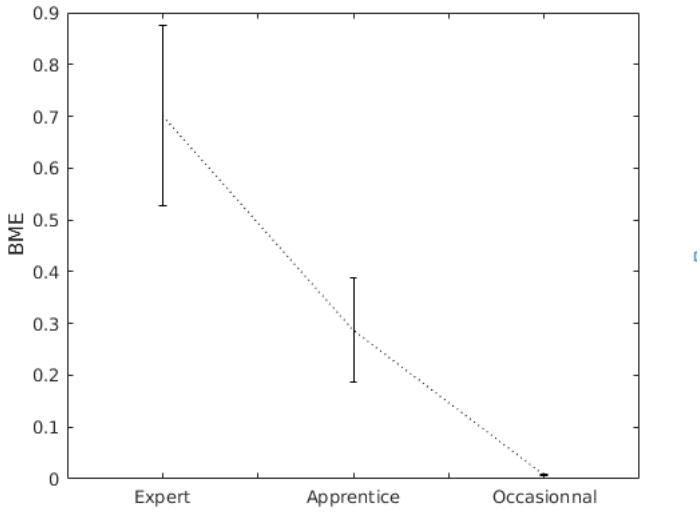


Figure 5.7: Error Bar BME

error bars. They are based on standard error (SE) bars and confidence intervals (CI). The mean of the data, with SE or CI error bars, allows us to have an indication of the region where we can expect the mean of population composing the dataset. The interval defines the values that are most plausible for the population used (Cumming et al., 2007).

We calculate the confidence interval with 95% for every class generated by the BME and Reputation. Results are described in Figures 5.7 and 5.8. Confidence intervals consist of a range of values (interval) that estimate the unknown intra-class parameters. We are 95% confident that the true value of Belief Measure of Expertise for experts is between $[0.3, 1]$.

Figure 5.8, we can see that the interval of experts' reputation is very large in $[2400, 35000]$. Thus a person whose reputation is equal to 2400 is an expert as an other contributor with 30000 reputation points. The fact that the reputation can not

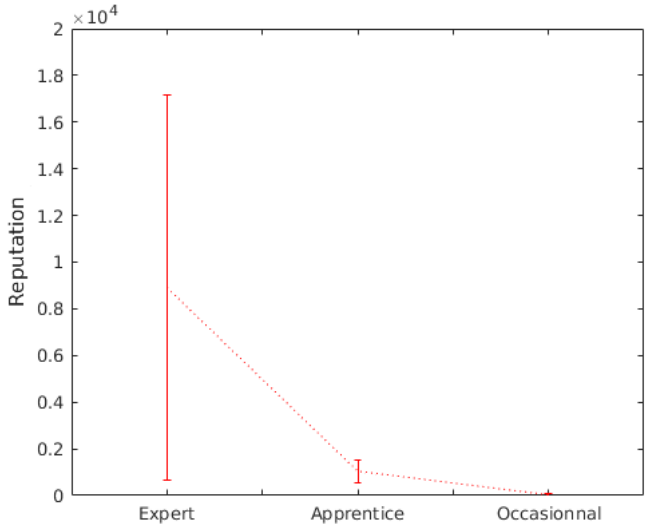


Figure 5.8: Error Bar Reputation

be enclosed, users may gain more and more points with no limitations, their expertise can not be similar to how the community member consider them based only on the reputation. As the values of the BME are in a limited interval of $[0, 1]$ and this measure takes into account both user's contributions and the time spent in the platform, this allows us to have an overview on how does a user evolve in the community. The BME is a general measure of expertise that does not take into account topical issues because of the popularity of some of them. This belief degree allows us to detect general expert users in the platform of Stack Overflow. The confidence intervals of BME described in Figure 5.7 shows that the standard deviation of all the three classes of users are relatively very small compared to those describing the reputation in Figure 5.8.

Hence, the mean value represents the good accuracy of the data. A small standard deviation on the bar means that the system is more reliable. If we have a larger standard deviation this means that the reputation system is less reliable

5.7 Human evaluation

After the evaluation of our classification model based on the theory of belief functions, we consider a human evaluation of the proposed approach. To do so, we randomly chose 500 users from the initial dataset. We label every user manually as Expert, Apprentice or Occasional. There are two types of correct decision resulting from the clustering process. The first type is called true positive (TP). Here, the classification assigns two similar objects to the same cluster. The second type is called True Negative (TN).

Table 5.2: Confusion matrix BME

	Occasionnals	Apprentices	Experts
Occasionnals	455	3	1
Apprentices	34	3	1
Experts	0	0	3

Table 5.3: Confusion matrix Reputation

	Occasionnals	Apprentices	Experts
Occasionnals	449	10	0
Apprentices	12	17	9
Experts	1	2	0

Here decision assigns two dissimilar objects to different clusters. Therefore, during the classification process there are two types of errors that can occur: a false positive (FP) decision assigns two dissimilar objects to the same cluster, while a false negative (FN) decision assigns two similar objects to different clusters.

5.7.1 Confusion matrices

For evaluation, we construct for each data set a confusion matrix (Provost and Kohavi, 1998) which contains information about actual classes and predicted ones. Tables 5.2, 5.3 and 5.4 represent respectively the confusion matrices of the BME, the reputation and the GMM. The confusion matrices allow the visualization of the performance of the classifiers. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class as defined by (Powers, 2011).
First of all, one can see that the behavior of our proposed approach based on the theory of belief functions, allows us to detect more experts than the two other approaches. However, we notice a miss-classification of the users belonging to the Apprentice class. Several apprentices have been considered as occasional users. We can see from the matrix that the BME trouble distinguishing between occasionals and apprentices. The Reputation and GMM show that they have better results for identifying the apprentices, especially the GMM. Therefore, the last approach seems to take the occasionals as apprentices. We notice a wrong detection of both experts and occasionals in Table 5.4.

The confusion matrices will allow us to measure the accuracy of the classification.

If we only consider the methods based on the reputation and GMM show us good classification for occasionals and apprentices. However, as our main purpose is to detect experts, the Belief Measure of Expertise present better results than the other methods. It seems that the GMM presents encouraging results when classifying apprentices. This can be witnessed by the results in Table 5.4.

Table 5.4: Confusion matrix GMM

	Occasionnals	Apprentices	Experts
Occasionnals	10	449	0
Apprentices	8	30	0
Experts	3	0	0

5.7.2 Performance evaluation

For evaluation, three metrics can be calculated: Precision (Prec), Recall (Rec) and F-measure (Fm). These metrics are from information retrieval domain and consist in comparing the obtained results with the expected labels (Fawcett, 2006).

- Precision is the proportion of correctly shared correspondences over the total number of found correspondences.

$$Prec = \frac{TP}{TP + FP}. \tag{5.9}$$

- Recall represents the proportion of correctly shared correspondences over the total number of referenced correspondences.

$$Rec = \frac{TP}{TP + FN}. \tag{5.10}$$

- F-measure represents the harmonic mean of precision and recall and is determined as:

$$F - measure = 2 * \frac{Prec * Rec}{Prec + Rec}. \tag{5.11}$$

The presented results in Figures 5.9.a, 5.9.b and 5.9.c describe respectively the values of the precision, recall and F-measure for the three clustering methods.

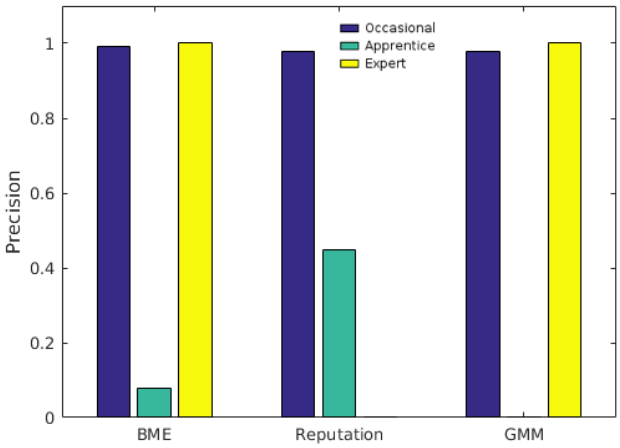
By comparing the obtained results using the different classification methods, it can be seen that the best results are obtained when using the Belief Measure of expertise based on the theory of belief functions.

In Figure 5.9.a we can see that the precision related to the classification of occasionals is very close to 1 for the three methods. However, for the Apprentices, the precision given by the reputation is more important than the one provided by our methods. Thus,

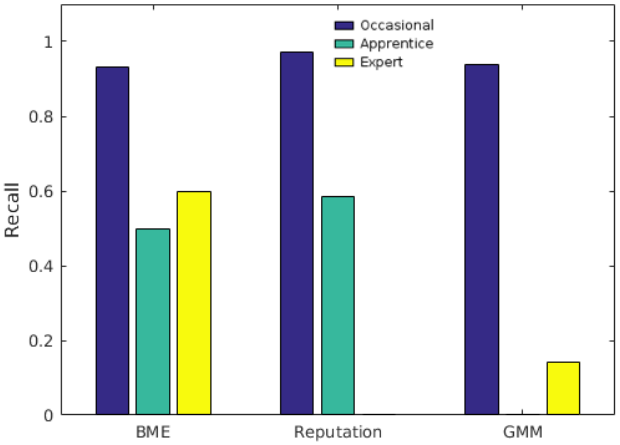
the GMM's precision for this class is null. For the experts BME and GMM present close values to 1. The precision of the detection is very good compared to the reputation. Therefore, the recall of experts with the GMM is really smaller than the BME.

5.8 Conclusion

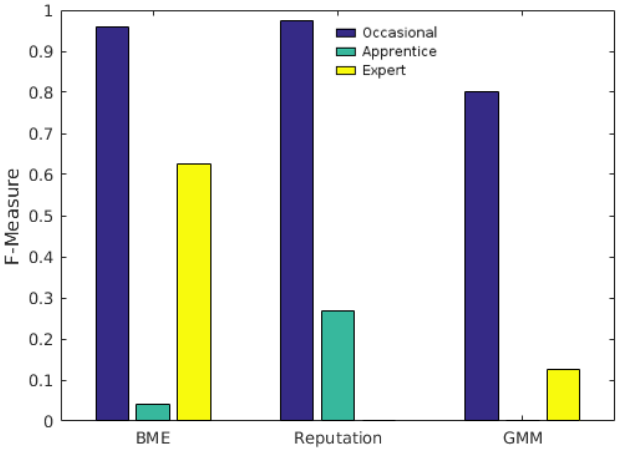
In this chapter, we propose a general measure of expertise for Stack Overflow based on the theory of belief functions. First, we presented the limitations of the reputation measure proposed by Stack Overflow and several other approaches. The main issue is that none of the other methods takes into account data imperfection. With the theory of belief functions, we can manage these imperfections and deal with them. Thus, we propose a belief measure expertise that considers users attributes and the time they spend on the platform. This metric is based on the combination of users' features which allows us to have a global overview about how they behave. The performance of this measure is evaluated on real data from Stack Overflow. The results of the BME are then compared with the reputation measure of Stack Overflow and a GMM.



a. Precision



b. Recall



c.F- Measure

Figure 5.9: Precision, Recall, F-Measure

6 Temporal Belief Measure of Expertise and Potential Experts detection

"Everything should be made as simple as possible, but **not simpler**"
Albert Einstein (1879 – 1955)

Contents

6.1 Introduction

In previous chapters, we were interested in clustering and analyzing the general behavior of users in the widely known question answering community Stack Overflow. We have proposed an attribute-based approach for experts detection in Stack Overflow. Experts are the "core" users within question answering community. Thus, new experts can be revealed whereas others may not be present for a long time. The members of a community evolve over time. They can belong to several class as long as they participate differently in the platform. Convinced that managing uncertainty in a experts detection process is an challenging task, we propose a temporal detection and analysis of expert users based on the theory of belief functions.
As Stack Overflow became very popular lately, this led the community to be opened to every person with an interest in programming. Thus massive activity may impact negatively the community. As stated by (Srba and Bielikova, 2016b), SO is failing its users.

The main contributions presented in this chapter are the following: first, we propose a global analysis of users' activity during several months of activity in the platform. We will examine which class is more concerned about posting questions or answers. Next, we will propose a study based on detecting potential experts during the first months after their registration and compare the results after a general classification based on the Belief Measure of Expertise. Later we propose a temporal analysis of three different classes of users and how do they behave for several months in the community. We will distinguish between three types of apprentices and experts that are present in the platform.
The sequel of this chapter is organized as follows: In the second section, we present the evaluation process for a temporal analysis of users. In section 6.2 we will present the data treatment and the proposed model for the temporal evaluation process. In section 6.3, we will describe the methodology in order to detect potential experts. We will compare the results with the general measure of expertise. Later in section 6.4 we will present a general overview of the temporal analysis of users in Stack Overflow. Finally in section 6.5 we will detail the changes that may occur to the every class of users. We will also present subcategories for the apprentices and experts according to their evolution over time in the community. Experiments made on real data sets provided by Stack Overflow are presented in this chapter to evaluate the performance of this model. This work is published in (Attiaoui et al., 2017b).

6.2 Evaluation process

In this section we will present the evaluation process for a temporal analysis of users in Stack Overflow. First we will start by describing how we process the data and divide it

into time buckets in order to have a specific overview of every user for every time snap. Next, we will present the model that will be followed in order to present a temporal analysis of experts and other users in the platform.

6.2.1 Data processing

The first step of the temporal analysis of users is to build the temporal series of number of questions, answers and accepted answers given by users during a period of time. To do this, we divide the periods of the data set into monthly buckets. The beginning of the first bucket is the time of the earliest question in the data set, noted t_0, and the end of the first bucket would be $t_0 + 30$ days. We work on data covering 15 months allowing us to have 15 time snaps for monthly buckets. The data used is from December 2013 to March 2015.

This bucketing system allows us to calculate the number of questions, answers, accepted answers and the votes generated by the posts during a defined period of time.

6.2.2 Model

In order to identify potential experts and make a temporal analysis of the three classes of users in the platform, we will use a modified version of the Belief Measure of expertise presented previously in section 5.3.2. Figure 6.1 describes the chart flow that will be used to do so.

At T_0, we calculate for each user the number of questions asked, answers posted, the scores generated and the number of accepted answers. Each value is transformed into mass functions using equations (5.1) to (5.5) as described in section 5.3. We do this process for data covering every 30 days. Thus for every period, we obtain for every user 5 features: 5 mass functions for the number of questions, number of answers, score of questions, score of answers and a mass function for the accepted answers. We combine these mass functions using the Demspter's combination rule presented in equation (3.13). Next we apply the pignistic probability and classify the user into Expert, Apprentice or Occasional for this specific time bucket.
At T_1, we use the results of the previous period and combine them with the mass functions of this actual period. After, we define the class of belonging. We maintain this combination and classification process for the entire dataset.

The combination process allows us to estimate the actual belief expertise called the Temporal Belief Measure of Expertise (noted TBME) for each user during a period is expressed by the following equation:

$$m^{T_1}(u_i) = {}^{\alpha^T}\left(m^i_{T_0} \oplus m^i_1 \oplus m^i_2 \oplus ... \oplus m^i_5\right) \tag{6.1}$$

$$TBME^{T_1}(u_i) = \ BetP(E) \tag{6.2}$$

where α^T is the discounting coefficient related to the time activity of a user. The value $\alpha_i^T = 1 - \frac{1}{nd_i}$ where nd_i is the number of days since the user first connected to the platform. The symbol $\oplus$ represents the operator of Dempster's combination.
TBME will be in the interval $[0, 1]$. This process of combination and classification for every time bucket allows to follow the progress of users monthly during a defined period of time. Furthermore, based on that, we can distinguish clearly the evolution of each user during their time activity within the community. Therefore, we can also detect potential experts on the onset of their participation.

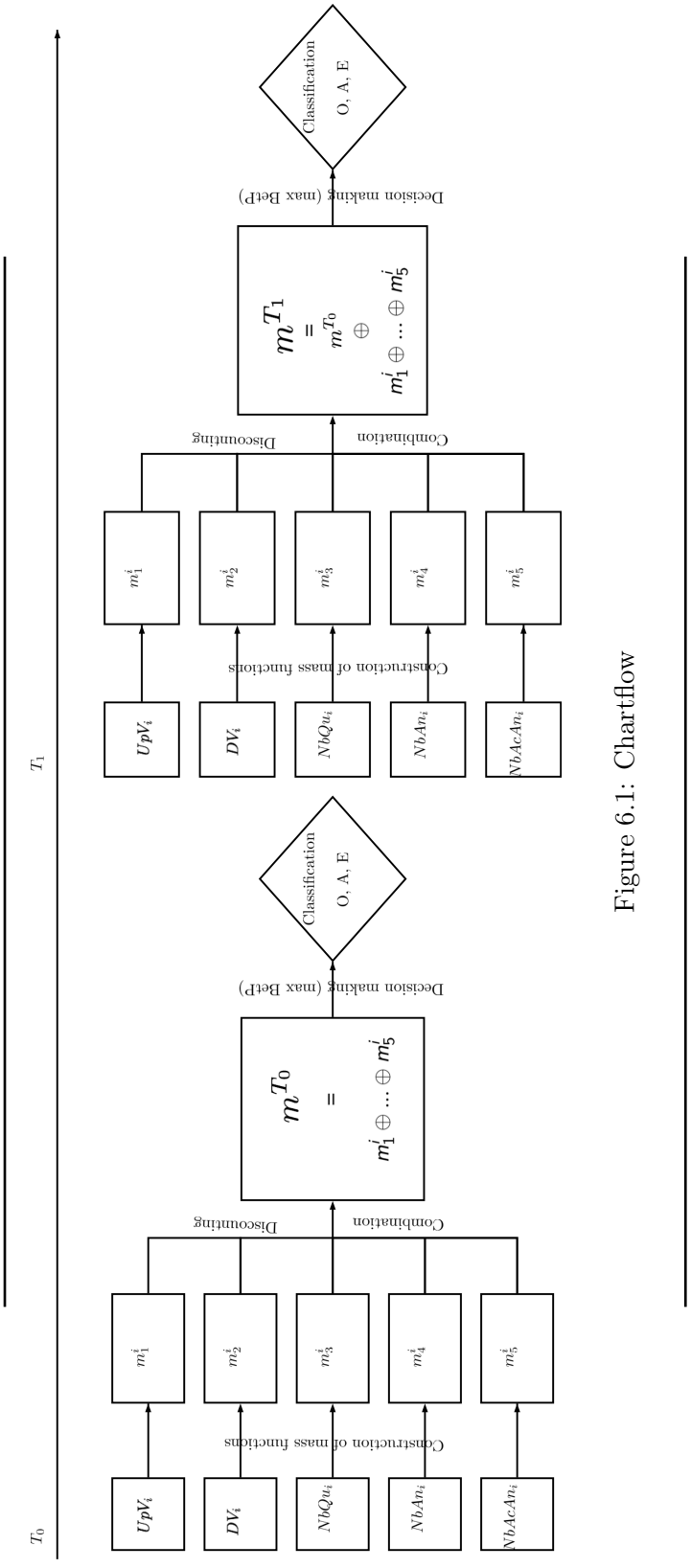


Figure 6.1: Chartflow

6.3 Detection of potential experts

(Yang and Wei, 2009) and (Nam et al., 2009) describe a potential expert as a person highly motivated to help the members of the community. Besides, this person should have to ability to answer questions correctly. (Pal et al., 2011) considered two major axis to evaluation: the motivation and the ability of a user. They used several indicators to detect these potential experts such as the number of contributions, their frequency, the votes gained by the posts, etc.

In this section we will try to identify these future experts and how they behave during the first months of their registration on the platform.

To do so we analyze the first n months of data for a user. We use all the features defined in Chapter 5 (score for answers and questions, number of posts, etc) and to build the model based on the theory of belief functions in order to identify potential experts. As (Pal et al., 2011) we will focus on how motivated these users are and how able they are to provide high quality content and contribute to the community.

6.3.1 Methodology

For this issue, we randomly choose 10.000 users in the platform. We calculate the number of their posts and obtained votes for 100 first days days after joining the community. We apply the TBME described in equation (6.2) and compare the results after several months of activity. We use the BME described in equation (5.6) for the final evaluation as it is a global measure of expertise.

6.3.2 Results

After several months of activity for some of the users, we notice that only 63% connected to the platform after 100 days after joining the community. We find several interesting observations. First of all, after 100 days of joining the platform, we can distinguish that 94.54% of users are classified as occasionals 4.252% as apprentices and $1,208$% as experts. Thus the percentage of experts is always the lowest compared to the other classes. Yet these users represent a small number to detect among the others. Few months later, we analyze the results of these users transformation within the community and identify their final classification.

- 85.90% of occasionnals detected remain newbies, they post few questions and answers. 1.51% of occasionals became apprentices. These users are trying to get involved with the other members of the community, but they still are not very active. Finally, 0.58% moved from occasionals to experts. These latter gained expertise over time, participate actively in the platform and became relevant users.

- 0.1004% of the users identified as apprentices during the first experiments remained in the same class. Some of them regressed to newbies with a percentage of 5.01%. Only very few 0.06% of them became experts.
- Only 0.02% keep being active and remain experts. Several users lose interest in the platform where 0.12% move to apprentices and 6.68% occasionals.

The experiment described here shows that modeling users and clustering them with the theory of belief functions and can be useful in finding high potential users within few months after joining the community. Thus identifying experts is already a hard task and detecting them 100 days after their registration in the platform is a challenging exercise.

6.4 General time analysis

In this section we will provide an overview analysis of how users behave during 15 months based on the number of questions asked and answers provided for the three classes of users.

Figures 6.2, 6.3 and 6.4, display respectively the distributions of the number of questions, answers and accepted answers over time. The values of the CDFs reach their maximum (which is equal to 1) only for experts and apprentices.
First we will start the analysis by figure 6.2 showing the CDF related to the mean of the number of questions posted by contributors over a period of time of several months. We notice that apprentices ask more questions than the other users. Considering the fact that these individuals are seeking information, and that they lack knowledge.

There is a common belief that experts do not ask questions in the community. However, here we can witness that experts do also questions over several time buckets. Most of the time, their CDF is lower than the apprentices', thus is relatively considerable compared to occasionals. This can be justified by the idea that experts can not know everything about anything: they are knowledgeable on some specific topics only. Moreover, they are known to post difficult questions that only other experts can answer.

We also notice that the values of occasionals' CDF values are smaller than the two other classes in general. These users ask from time to time questions only when they need information has not already been discussed on the platform.

The CDF related to the mean of the number of answers is represented in figure 6.3. We notice the same phenomenon described for the CDF of questions. At the beginning both experts and apprentices have almost similar values. However, over time, experts are less and less present within the community. They do not provide as much answers as the apprentices. This can be explained by the desire of apprentices to gain popularity and expertise. As some of them are able to provide correct answers quickly to easy questions allows them to be very active in the platform. However, some of the apprentices try to

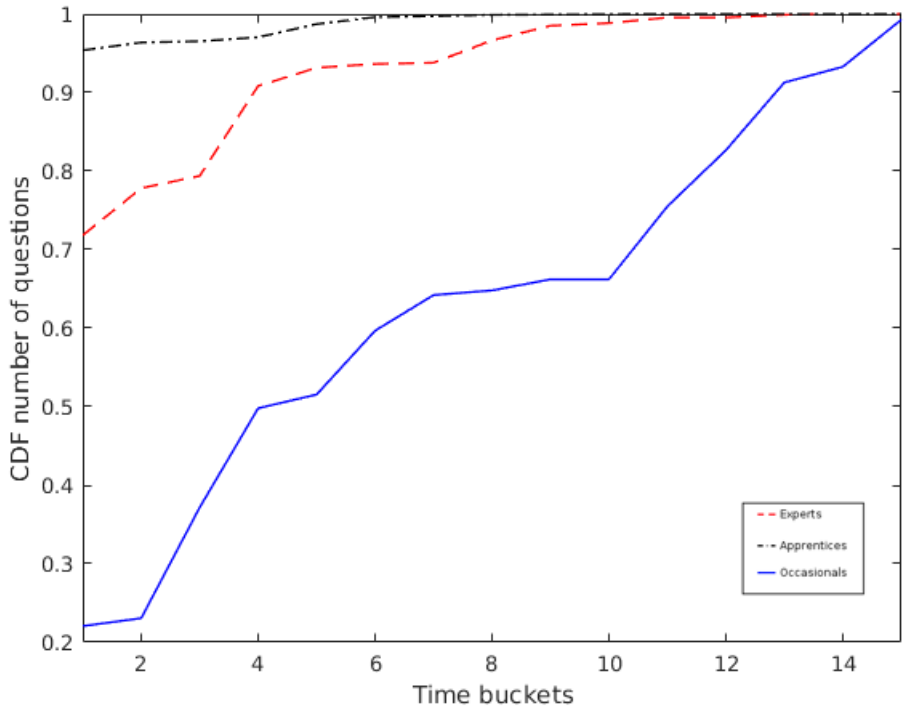


Figure 6.2: CDF number of Questions

provide a lot of contributions because they are motivated by gaining reputation points in Stack Overflow, sometimes neglecting the quality of their posts. The fact that users post less and less questions over time like shown in figure 6.2 may discourage experts to share their knowledge on the platform. This can cause the decrease of their interest on posting helpful answers.

The number of accepted answers is a very important indicator on how to evaluate the expertise of a user in Stack Overflow. The CDF of the number of best answers provided by each class of users is presented in figure 6.4. We can see that experts are gaining the highest values of accepted answers. For the apprentices, they seem to be having knowledge as their mean CDF is close to the values experts. We see some of them are becoming future experts. They post a lot of answers that are chosen as the best. The more time they spend on the community the more expertise they have.

6.5 Analysis of users over time

In this section we provide an analysis of the activity of the users during the 15 months of the dataset.

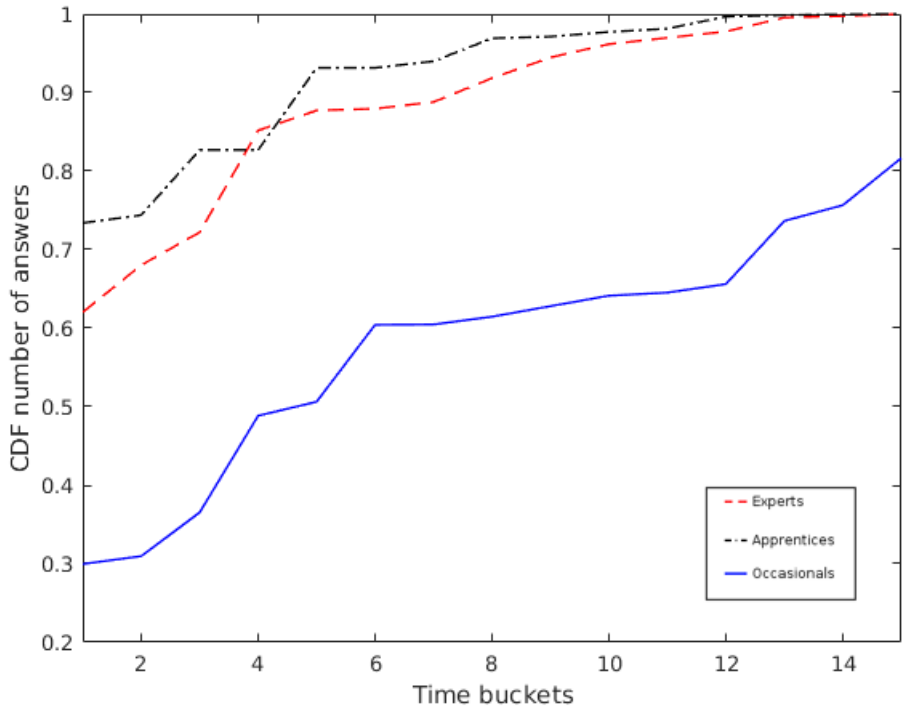


Figure 6.3: CDF number of Answers

6.5.1 Evolution of number of users

As described before, we classify users according to the belief expertise measure displayed in equation (5.6) for every time bucket. We randomly choose n users from the big dataset and we obtain the results presented in figure 6.5.

For each time bucket we obtain the percentage of users according to every class of users. First of all, we notice that the number of Occasionals is always has the highest representation of users in the platform. After that, proportionally to the number of newbies, apprentices are not that numerous. However, we witness that their number does not change a lot over the months. Their percentage seems to be almost the same through the time.

Finally, for the experts, we find that their number fluctuates over the period of time described in the dataset. For the last time buckets, they become more and more scarce. The community may risk high-potential users leaving because of the lack of recognition regarding their efforts by other contributors. The number of apprentices staying the same can be explained by two phenomena. The first one is that some occasionnals became apprentices, or some experts are loosing interest in the community and are

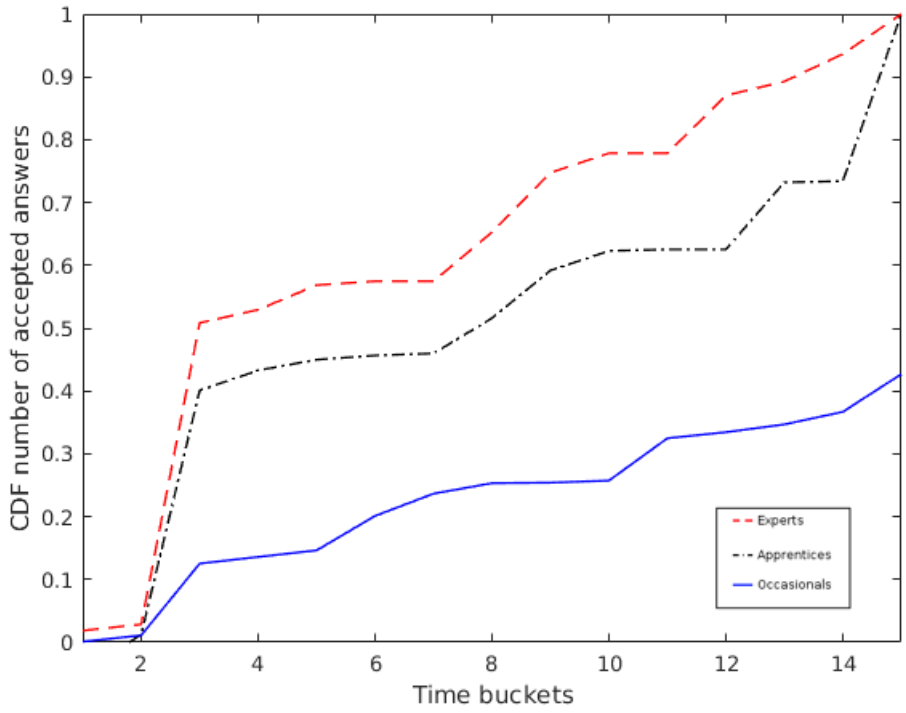


Figure 6.4: CDF number of Accepted Answers

posting less.

6.5.2 Evolution of Occasionals

Figures 6.6 and 6.7 describe respectively some evolution of Occasionals and their TBME over time. As shown before, the occasionals represent the most numerous class of persons registered in the platform.
Besides the fact that the majority of the occasionals remain in the same class for a long time, we may come across some exceptions. We can find users that move from a class to another according to their motivation and the intensity of their participation in the community. Figure 6.6.a both represents the evolution of an occasionnal to apprentice. While figure 6.7.a describes the values of the Temporal BME over the months.

Figure 6.6.b describes a user that was a newbie for the first months of the dataset. Later, this person become an apprentice, participates with moderation in the platform. Their activity during this time allowed him to reach a value over 0.3 for TBME. Therefore, we can see that some individuals are classified as newbies in the beginning and change from a class to another as shown in figures 6.6.c and 6.7.c. This person moves

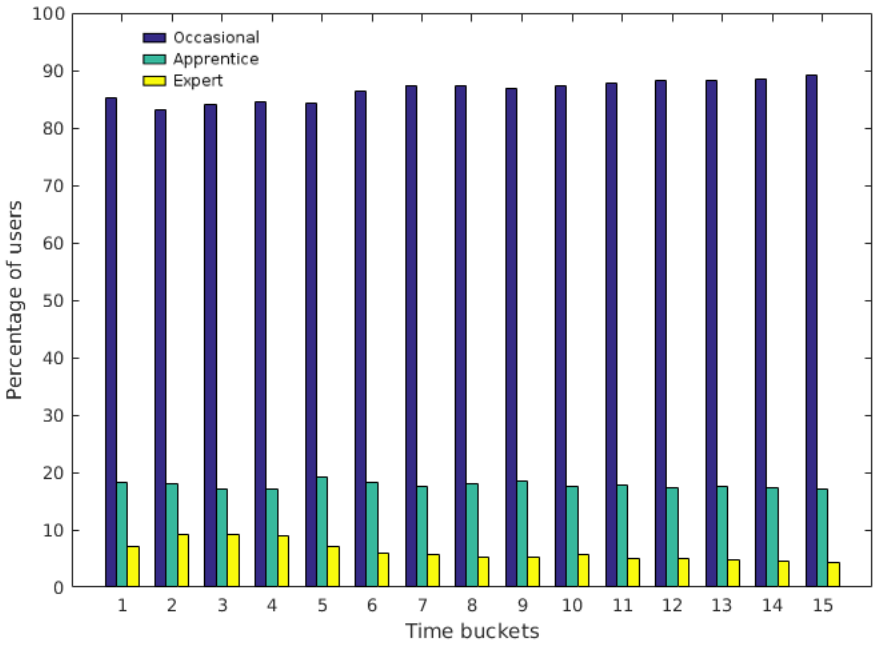
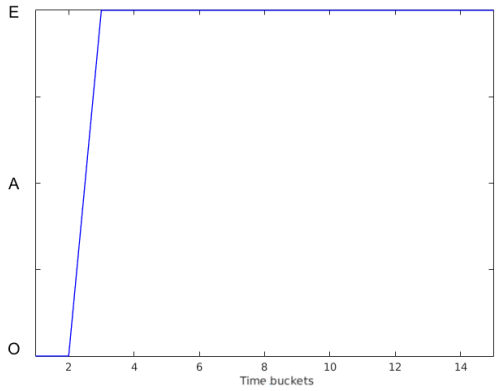


Figure 6.5: Evolution of the percentage of Occasionals, Apprentices Experts per time bucket

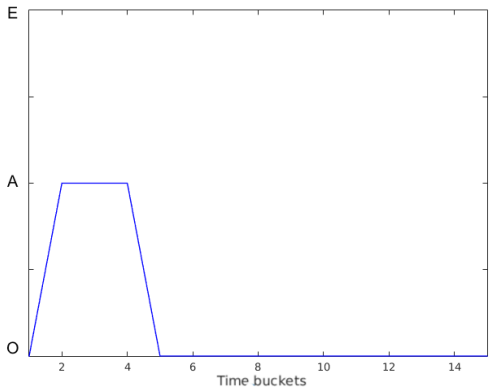
gradually from occasional to apprentice to expert after several months spent in the community. This means that this user have gained knowledge when joining the platform and become helpful by providing high quality content. This is described with the value of the TBME in figure 6.7.c. However, we can see that at the end of the time buckets this user's measure of expertise decreased for the last months but still considered as an expert.

6.5.3 Evolution of Apprentices

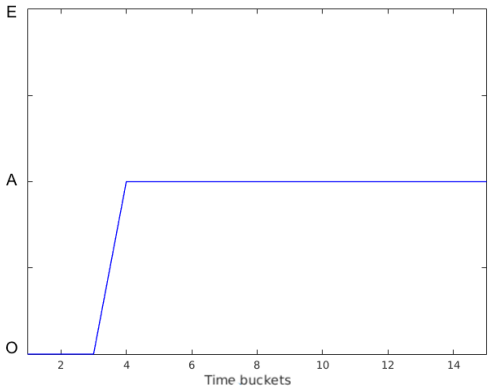
Apprentices are the very active users in the platform. Their main motivation is to gain knowledge and reputation points. Reputation is the main indicator about popularity of a user in his community. Figures 6.8 and 6.9 describe respectively some evolution of Apprentices and their TBME over time. Some of these users change from a class to another based on the quantity and quality of their contributions. We can see that a learner can quickly become an expert as show in figure 6.8.a after few months. These individuals are very motivated members with knowledge that only need a short amount of time to impose themselves as important users within their community. However, we



a. Occasional to Expert



b . Occasional to Apprentice to Occasional



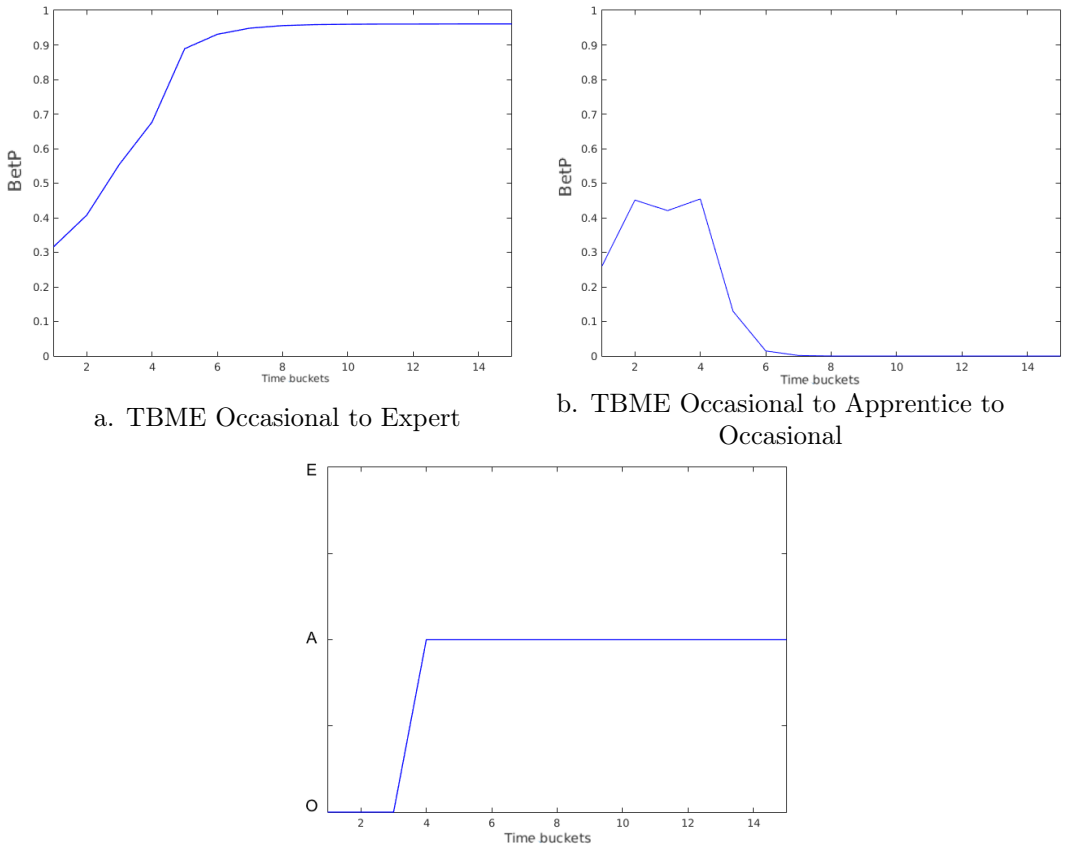
c. Occasional to Apprentice

Figure 6.6: Evolution of Occasionals

can see other apprentice evolve differently.

Figure 6.8.a describes the possibility of an apprentice becoming an expert. The figure shows that the expertise is revealed very quickly. Actually this kind of users are the ones who gain knowledge by participating to the community. Over the time spent, they learn about any topic that interests them and gradually become knowledge and able to provide high quality content. This is mirrored by the values of their TBME that increases as shown in figure 6.9.a.

As in figure 6.8.b, this user became and expert and gradually over time he posted less questions and answers, making his TMBE decreasing and making him an occasional. The TBME fluctuates over the months where later after the 5^{th} month the TBME has a null value. This is explained by the fact that this user does not post anymore in the platform.



a. TBME Occasional to Expert

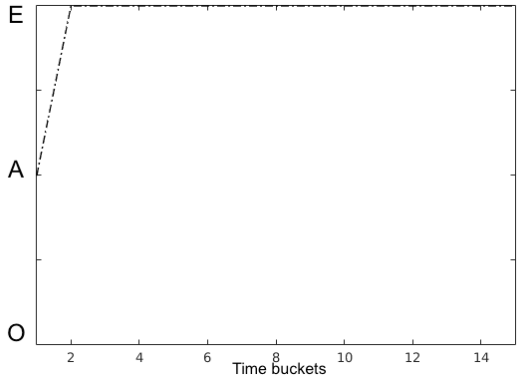
b. TBME Occasional to Apprentice to Occasional

c. TBME Occasional to Apprentice

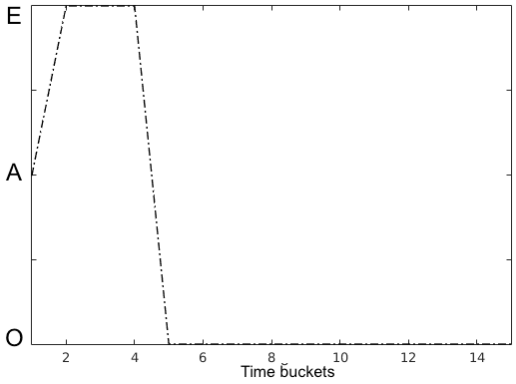
Figure 6.7: Evolution of TBME Occasionals

Consequently we can determine three main types of apprentices:

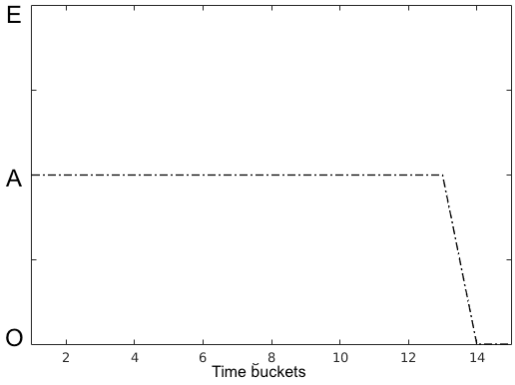
- Quick Apprentices: these users show their expertise after a short time spent in the community.
- Slow Apprentices: these users gain knowledge after posting a lot of questions and answers in the community. They need several months to become experts and share with the community.
- Uninterested Apprentices: these users were motivated in the beginning and they lose interest over time, due to the negative votes they may gain by posting low quality content and the competition they meet in the platform. This may lead them to leave to the community and/or become occasionals.



a. Apprentice to Expert



b. Apprentice to Occasional to Expert



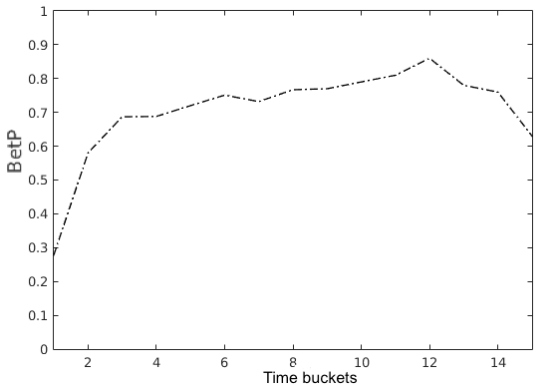
c. Apprentice to Occasional

Figure 6.8: Evolution of Occasionals

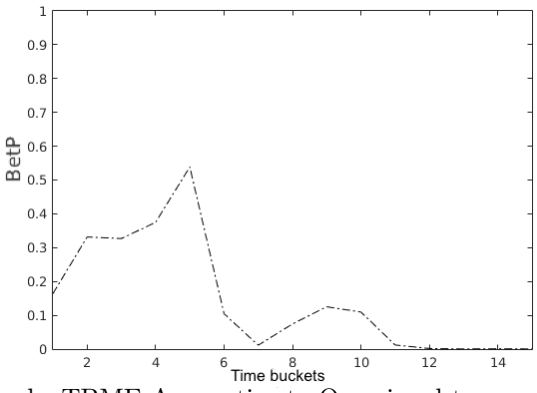
6.5.4 Evolution of Experts

As experts are the most important answers providers, the quality of exchanges in the platform directly depends on their motivation and their contributions. During the 15 months of our evaluation, we found that some experts remain in the same class for a long time. These users provide a lot of answers and accepted answers making their TBME always high and close to 1. These users maintain their activity in the platform, they are very active and concerned by helping others who are persons seeking for information. As decribed by (Pal et al., 2012a) we find the same results when categorizing experts. Thus, figures 6.10 and 6.11 describe respectively some evolution of Expertise and their TBME over 15 months. Here we can distinguish between 3 possible final behaviors of experts in Stack Overflow:

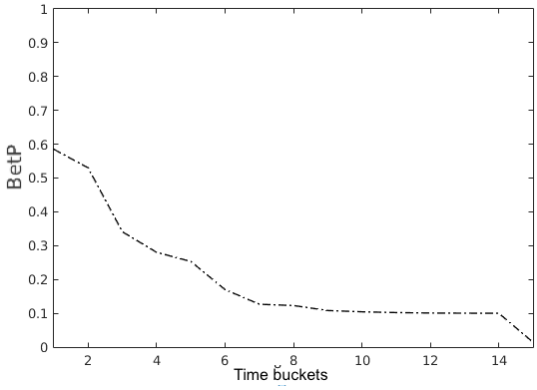
- Consistently active experts: for this type of experts, their motivation and activity



a. TBME Apprentice to Expert



b. TBME Apprentice to Occasional to Expert



c. TBME Apprentice to Occasional

Figure 6.9: Evolution of TBME of Apprentices

are always at peak. They participate at same frequency during the several months of the experiments.

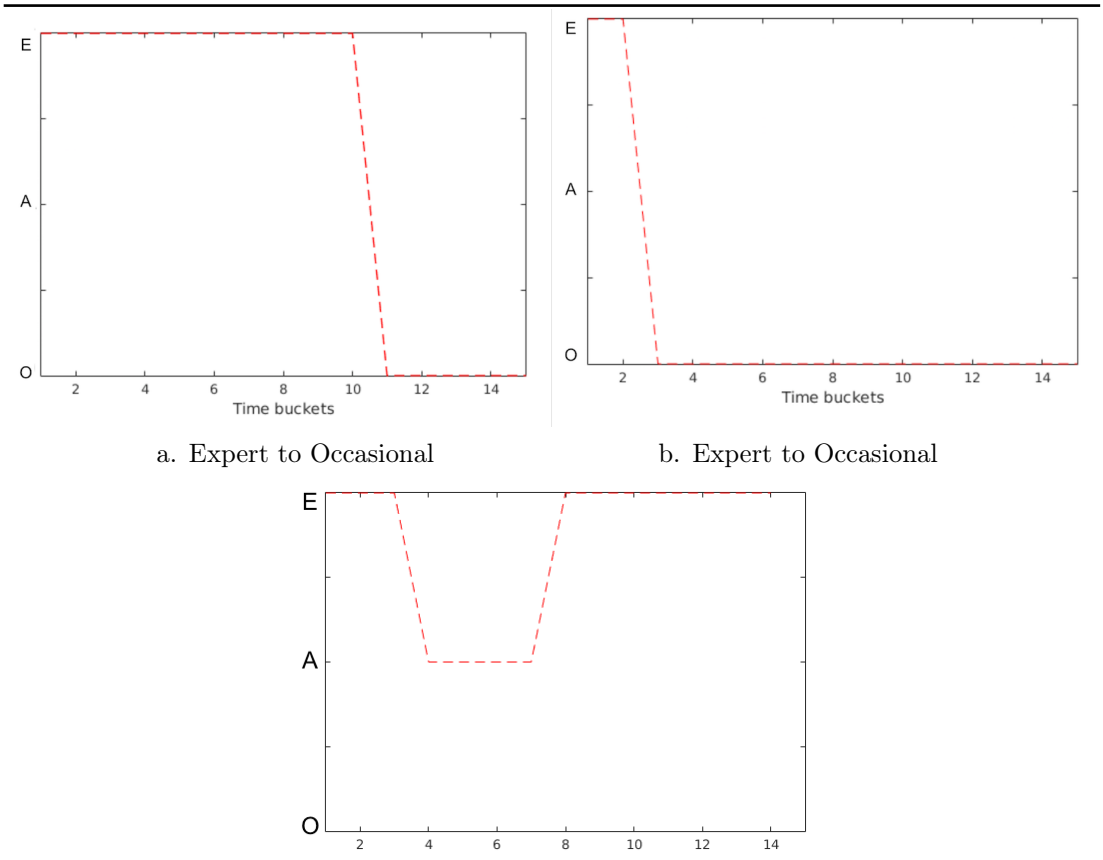
- Active but later passive experts: for this type of experts, the community was attractive at the beginning however gradually lost their motivation and interest on posting questions or answers.
- Passive but later active experts: for this type of experts, they observe before they post, they analyze the community. They are not very talkative, they do not post a lot at the beginning but later, they prove their ability to help the other members of the community by participating more and more.

These behavioral changes are confirmed by figure 6.10 and their temporal measures of expertise in the figure 6.11. Figure 6.10.a shows how a user can evoluate from an

expert to an occasional after several months of activity in the platform. As stated before in figure 6.5 experts are loosing interest and are leaving the community as their number decreases over time

Figure 6.10.b shows how an expert becomes later passive throw the time-line of our experiments. Their activity gradually decreases by becoming apprentices posting less and then to occasionals, participating from time to time or not at all. This is correlated by the value of their TBME as shown in figure 6.11.b.

Figure 6.10.c represents expert's classification over 4 months. This behavior changed and he became a learner. This means that this user became less active for few months and become motivated again by participating in the platform during the 7^{th} time bucket. This allowed him to re-become an expert.

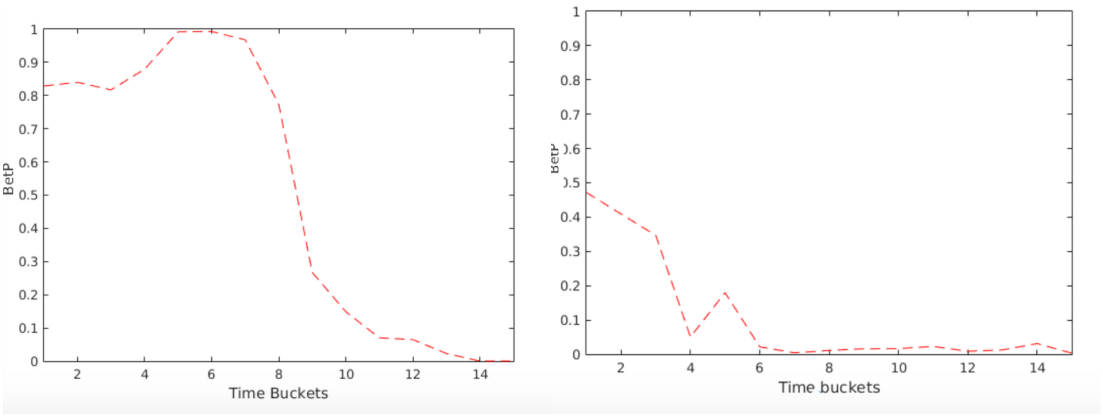


a. Expert to Occasional

b. Expert to Occasional

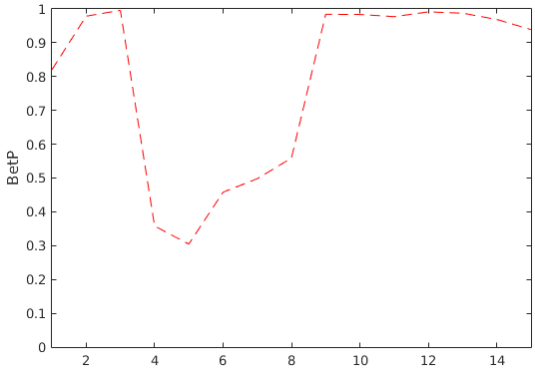
c.Expert to Apprentice to Expert

Figure 6.10: Evolution of Occasionals



a. TBME Expert to Occasional

b. TBME Expert to Occasional



c. TBME Expert to Apprentice to Expert

Figure 6.11: Evolution of TBME Experts

Mainly active experts are the most influential users in the platform. They are able to manage the difficult questions by responding quickly with very helpful answers. Some of them stay always as experts and some stay always as occasionals based on their initial classification. However, we can witness the evolution of some members over the time spent within the community.

We can see that contributors may evolve during their time activity in the platform from occasional to apprentice to expert. Thus, we may notice that some users interrupt their contribution for some months and then restart posting. For some other users. Therefore, we can find users that can be experts for a period and then start posting less and less until leaving the community becoming occasionals.

6.6 Conclusion

This chapter is focused at two major issues: first on identifying three classes of users on Stack Overflow: Occasionals, Apprentices and Experts. Then, detecting potential experts on their early time of activity. The strength of the proposed model is that it could be applied to any topic in the platform. Based on a belief model of the users' behavior, we calculated the general degree of expertise called the TBME. This measure takes into account the combination of all the masses that describe a user during a defined period of time of activity on the web site. Once the expertise measure calculated for each time bucket, it allows us to have an overview of the users' behavior. Potential experts can be detected since the early few months of their entrance to the community.

The next chapter is dedicated to present some conclusions of this thesis.

7
Conclusion and perspectives

7.1 Conclusion

Question Answering Communities became gradually important by changing the way we seek information. People count on other user's expertise to help them find a solution to any issue. All they have to do is to post a question in the right community. These online forums offer a valuable archive of information at several levels: information about users, their personal and professional interests, and finally their ability to provide helpful answers or post well expressed questions. In Stack Overflow like any other online community, every member is characterized by some features: scores, votes, posts, comments, etc.

We provide a data analysis on the data downloaded from Stack Overflow. We performed a principle component analysis and a mixed classification. For the latter, we used both partitioning and hierarchical clustering techniques. We identified the most important features related to online community users. In this thesis, based on these attributes we were able to propose a new general measure of expertise founded on the theory of belief functions. This expertise measure allows us to estimate the degree of knowledge of users in question answering communities. The theory of belief functions is used as a strong tool in order to model and combine information related to every individual.

We have identified three classes of users: Expert, Apprentice and Occasional. Every class is characterized by its own behavior and how active are the users. Their activity and their motivations differ from a class to another. Experts aim to help other members and share their knowledge with them by providing high quality answers. Apprentices are more motivated by gaining reputation points and notoriety in their community. While occasionals post questions from time to time in order to obtain responses. We compared the results of our belief clustering approach to the reputation-based system of Stack Overflow and a Gaussian Mixture Model. We have shown that our classification presents better partitions on some axis. Actually our belief model can be considered as an attribute based approach. We only take into consideration users features in the platform such as the number of questions, answers, accepted answers, etc. We compared our approach to the reputation based method proposed by Stack Overflow. For a lot of users reputation is the reflection of their expertise. Actually, this measurement only describes how popular a user is in the community.

The majority of users in online communities are occasionals, making experts' number

very small. Consequently, identifying them can be considered as a challenging task. Once this step of experts identification achieved, we focused on finding potential experts few months after joining the community.
Last but not least, we performed a temporal analysis on users in Stack Overflow. To do so, we divided our dataset into times series. For each one we calculate the number of posts and scores generated on a monthly basis for every user during 15 time buckets.

We evaluated the evolution of all types of users during their time activity in the platform. Some users remain in their initial classes: always an occasional or always an experts. Where some other change their activity leading to a change of their classification. Hence, we distinguished between three types of experts: constantly active, from active to passive or from passive to active. Thus, for the case for apprentices, we identified: quick, slow or uninterested. This time analysis provides a general overview on how can users evolve in the community over time.
Lately Stack Overflow have known lack of interest from online users. This can be proven by two major posts in the network. In the first post users are wondering *why is Stack Overflow is becoming negative*?[21]. For the second one, contributors notices that the quality of posts decreased lately[22].
Actually, the website is very competitive and encouraging individual efforts (Matei et al., 2017). Users are trying to find controversial ways to improve their reputation scores. It seems that new users are pushing experts and learners away from the platform making the most expert contributors leaving the community. This is directly reflect by a lack in the quality of the posts in Stack Overflow.
Actually we came across these conclusions in the last chapter where we noticed that the number of experts decreased over time buckets, where occasionals are becoming more numerous in the platform.
It seems that Stack Overflow is a victim of its own success. As the platform became very popular and opened to every one, this situation led to the decline of the posts quality. Apparently, the questions asked do not interest expert users. Occasionals ask many questions without trying to find a solution in the archives or in the web. This makes a lot of redundant posts that have already been treated. As experts are more interested on difficult and attractive questions, they have to deal with low quality questions making them less and less active in the community. Therefore, in Stack Overflow, we can also cross a lot of apprentices who are concerned only by gaining as much reputation points as possible without caring about the quality of their posts. In order to gather reputation, these users publish a lot of answers regardless of their lack of expertise.
Among the other reasons of this phenomenon is the fact that a lot of users do not respect the guidelines governing the community or the fact that some questions treat with software that are not inherently technical or not famous are regularly down voted/closed.

[21] https://meta.stackoverflow.com/questions/251758/why-is-stack-overflow-so-negative-of-late
[22] https://meta.stackoverflow.com/questions/252506/question-quality-is-dropping-on-stack-overflow

7.2 Perspectives

In this thesis, we attained some new findings and presented interesting results using the theory of belief function. However, many other improvements have yet to be achieved and more work and research can be conducted . In the following, we introduce some perspectives for future works:

- Topical belief measure of expertise: In this thesis, we only considered the general measure of expertise. It would be interesting to have a topical measure of expertise and a way to measure a distance between topics. This would allow us to have a detailed overview of user's expertise from a topic to an other. Then, this topical belief measure of expertise will allow us to study the evolution of expertise between different expert users and to understand how do users in a topic learn from others. If a user has a average global measure of expertise, with a high number of questions downvoted in "Ruby" and at the same time he is very knowledge in an other topic such as "Python" with several accepted answers. It would be interesting to differentiate between these topics by using a topical distance that discounts the global expertise. An other idea, is to evaluate his evolution on "Ruby" or an other programming tool over the time spent in the community.

- Expertise propagation can be an interesting area to focus on. How can expert learn from an other experts. For this matter, we would only take into consideration expert users. The analysis of the behavior of a small community inside the global community can lead to an attractive research. Thus, we can be able to investigate how do experts perform when they are dealing only with knowledgeable peers. Thereby, we can explore to propagation of expertise in a global and topical matter.

- We would also like to explore some other data provided by other question answering communities such as Quora[23] or WikiAnswers[24]. It would be interesting to test the results of this measure of expertise in other collaborative web site.

- Inter-Community Expertise can be an other idea is to investigate. Considering a person being member of several platforms, what makes him/her prefer a community to another? How would he/she behave in different online forums? We can also measure an inter-communities expertise.

- Optimization of the parameters of BME and TBME. We uses several parameters during the construction of the mass functions and their combination. However, we defined these values manually. It would be interesting to investigate an optimization research field like genetic algorithms in order to provide the optimal value for each parameter.

[23] https://www.quora.com/
[24] https://www.wikianswers.com/

A
Indices of the clustering' quality

We presented clustering validity checking approaches based on internal and external criteria. For the external criteria, the indices measure if a clustering is similar to a model partition P. It is equivalent to have a labeled dataset (ground thruth).

The internal criteria allows to measure some properties that are expected in a good clustering like the how compact are the groups and are they well separated. These indices are based on the attributes values measuring the properties of a good clustering. The criteria also take into consideration the statistical properties of the attributes of the model values distribution and distances distribution

In this section we will only focus on the internal criteria. We present some indices used for the evaluation of the quality of the clustering.

For C_i the class of index i between N_c different classes and n_i the number of elements of C_i we study the following indices:

- Intra class inertia (Lebart et al., 1980): measures the homogeneity between the objects within the same cluster. A good clustering has small values.

$$Intra = \frac{1}{n}\sum_{i=1}^{k} \quad n_i \quad d(c_i, c)^2 \tag{A.1}$$

- Inter class inertia (Lebart et al., 1980): measure the degree of heterogeneity or separability between clusters. A good clustering has high values.

$$Inter = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} d(c, c_G)^2 \tag{A.2}$$

  with c_G the gravity center of the class.

- Silhouette (Rousseeuw, 1987): this criterion focuses on the objects within a cluster C_i. It measures if every object have been well classified.

$$-1 \leq S(i) = \frac{b(i) - a(i)}{max\left\{a(i), b(i)\right\}} \leq 1 \tag{A.3}$$

with $a(i)$ the mean distance between an object and its peers in the same class they belong to and $b(i)$ the mean distance between an object and the other objects of the closest class. When the value of S is close to 1 this means that the object is in the right cluster.

- Davies-Bouldin (DB) (Davies and Bouldin, 1979): this criterion treats every cluster individually. Its measures how similar a class is to the closest class. The best partition has to be minimizing the mean value calculated for every cluster.

$$DB = \frac{1}{n}\sum_{i=1}^{N_c} max\{\frac{I(C_i) + I(C_j)}{I(C_i, C_j)}\} \tag{A.4}$$

where $I(C_i)$ represents the mean of the distances between the objects of a cluster and its center. The sum $I(C_i) + I(C_j)$ represents the distance between the centers of two clusters.

- Calinski-Harabasz (CH) (Caliński and Harabasz, 1974): this criterion weighs the intra cluster variance by the number of classes. Its maximization is the optimal partition.

$$CH = \frac{SS_B}{SS_W} \times \frac{(N - C)}{(C - 1)} \tag{A.5}$$

where SSB is the overall between-cluster variance, SSW is the overall within-cluster variance.

- Dunn (Dunn, 1974): The Dunn's index measures compactness (maximum distance in between data points of clusters) and clusters separation (minimum distance between clusters). The maximum value of the index represents the right partitioning. The goal is therefore to maximize the inter-cluster distance while minimizing the intra-cluster distance. Dunn's index is described according the following equation:

$$Dunn = min_i \left\{ min(\frac{d(_{C_i,C_j})}{max(intra(C_i))}) \right\} \tag{A.6}$$

where $d_{(C_i,C_j)}$ is the distance between cluster C_i and C_j and $intra(C_i)$ the intra-cluster function of the cluster.

- Root-Square (RS): also called coefficient of determination. It measures the degree of difference between clusters. When it is close to 0 this means that the predictive model is weak. Otherwise, when the RS near 1 we have a strong clustering. The best partition has to be close to 1. RS is expressed by the following equation:

$$RS = \frac{\sum_{i=1}^{N_c}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N_c}(y_i - \bar{y}_i)} \tag{A.7}$$

where y_i is the observed value, $\bar{y}$ as its mean, and $\hat{y}$ as the fitted value.

After defining some of methods proposed in the literature for the evaluation of the quality of the clustering, in Table A.1 we summarize the rules to follow in order to determine the best partition.

Table A.1: Method to determine the best partition

Index	Rule
Intra	small
Inter	high
Silhouette	high
Davies Bouldin	small
Calinski-Harabasz	high
Dunn	high
RS	high
RMS	small
RS Error	small

B
Publications

The proposed approaches have been the subject of four publications. Three are published in international conferences whereas the reminder have been published in a national workshop:

1. Imen Ouled Dlala, Dorra Attiaoui, Arnaud Martin, Boutheina Ben Yaghlane: Trolls Identification within an Uncertain Framework. International Conference on Tools with Artificial Intelligence, ICTAI, Limassol, Cyprus, 2014: 1011-1015

2. Dorra, Attiaoui, Arnaud Martin, Boutheina Ben Yaghlane: "Détection des experts dans un cadre incertain" in Plate-forme Intelligence Artificielle, 2015, Rennes, France

3. Dorra Attiaoui, Arnaud Martin, Boutheina Ben Yaghlane: Belief Measure of Expertise: Case Study Stack Overflow. Big Data Analytics and Knowledge Discovery - 19th International Conference, DaWaK 2017, Lyon, France, 368-382

4. Dorra Attiaoui, Arnaud Martin, Boutheina Ben Yaghlane: Belief Temporal Analysis of Expert Users: Case Study Stack Overflow. 21st International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, KES 2017, Marseille

Bibliography

Ackerman, M. and McDonald, D. (1996). AnswerGarden 2: Merging organizational memory with collaborative help. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work*, pages 97–105.

Adamic, L. A., Zhang, J., Bakshy, E., and Ackerman, M. S. (2008). Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, pages 665–674. ACM.

Aslay, C., O'Hare, N., Aiello, L. M., and Jaimes, A. (2013). Competition-based networks for expert finding. In Jones, G. J. F., Sheridan, P., Kelly, D., de Rijke, M., and Sakai, T., editors, *SIGIR*, pages 1033–1036. ACM.

Attiaoui, D., Martin, A., and Ben Yaghlane, B. (2015). Détection des experts dans un cadre incertain. In *Plate-forme Intelligence Artificielle*.

Attiaoui, D., Martin, A., and Ben Yaghlane, B. (2017a). Belief measure of expertise for experts detection in question answering communities: case study stack overflow. In *21st International Conference on Knowledge-Based and Intelligent Information Engineering Systems-KES*.

Attiaoui, D., Martin, A., and Ben Yaghlane, B. (2017b). Belief temporal analysis of expert users: case study stack overflow. In *19th International Conference on Big Data Analytics and Knowledge Discovery - DaWaK*.

Balog, K. and de Rijke, M. (2009). Combining candidate and document models for expert search. In Voorhees, E. M. and Buckland, L. P., editors, *TREC*, volume Special Publication 500-277. National Institute of Standards and Technology (NIST).

Bouguessa, M., Dumoulin, B., and Wang, S. (2008). Identifying authoritative actors in question-answering forums: the case of yahoo! answers. In Li, Y., Liu, B., and Sarawagi, S., editors, *KDD*, pages 866–874. ACM.

Bouguessa, M. and Romdhane, L. B. (2015). Identifying authorities in online communities. *ACM TIST*, 6(3):30.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics Simulation and Computation*, 3(1):1–27.

Chan, J., Hayes, C., and Daly, E. (2010). Decomposing discussion forums using common user roles. *Proceedings of the WebSci10: Extending the Frontiers of Society OnLine*.

Chase, W., Simon, H. A., and Chase, W. (1973). *The mind's eye in chess*. Academic Press, New York.

Chase, W. G. and Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4:55–81.

Chebbah, M., Martin, A., and Ben Yaghlane, B. (2015). Combining partially independent belief functions. *Decision Support Systems*, 73:37–46.

Chen, Y., Ho, T.-H., and Kim, Y.-M. (2010). Knowledge market design: A field experiment at google answers. *Journal of Public Economic Theory*, 12(4):641–664.

Chirag Shah, S. O. and Oh, J. S. (2009). Research agenda for social q a. volume 31 of *library and information science research journal*, page 205âĂŞ209. Springer.

Choi, E., Kitzie, V., and Shah, C. (2012). Developing a typology of online q&a models and recommending the right model for each question type. *Proceedings of the Association for Information Science and Technology*, 49(1):1–4.

Cleuziou, G. (2004). *Une méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information. (A Clustering method for rules learning and information retrieval).* PhD thesis, University of Orléans, France.

Cumming, G., Fidler, F., and Vaux, D. L. (2007). Error bars in experimental biology. *The Journal of cell biology*, 177(1):7–11.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1(2):224–227.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. In *Classic Works of the Dempster-Shafer Theory of Belief Functions*, volume 219.

Denoeux, T. (1995). A k-nearest neighbor classification rule based on dempster-shafer theory. *IEEE Trans. Systems, Man, and Cybernetics*, 25(5):804–813.

Dlala, I. O., Attiaoui, D., Martin, A., and Ben Yaghlane, B. (2015). Trolls identification within an uncertain framework. abs/1501.05272.

Dubois, D. and Prade, H. (1988). Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence*, 4:244–264.

Dubois, D. and Prade, H. (2015). Possibility theory and its applications: Where do we stand? In *Handbook of Computational Intelligence*, pages 31–60. Springer.

Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4:95–104.

Ericsson, K. A. (2006). The cambridge handbook of expertise and expert performance.

Essaid, A., Martin, A., Smits, G., and Ben Yaghlane, B. (2014). A distance-based decision in the credal level. *International Conference on Artificial Intelligence and Symbolic Computation AISC*.

Estivill-Castro, V. and Yang, J. (2000). Fast and robust general purpose clustering algorithms. In *PRICAI*, pages 208–218.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). Advances in knowledge discovery and data mining.

Forestier, M., Stavrianou, A., Velcin, J., and Zighed, D. A. (2012). Roles in social networks: Methodologies and research issues. *Web Intelligence and Agent Systems*, 10(1):117–133.

Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588.

Furlan, B., Nikolic, B., and Milutinovic, V. (2013). A survey and evaluation of state-of-the-art intelligent question routing systems. *Int. J. Intell. Syst.*, 28(7):686–708.

Furtado, A., Andrade, N., Oliveira, N., and Brasileiro, F. V. (2013). Contributor profiles, their dynamics, and their importance in five q&a sites. pages 1237–1252. ACM.

Gazan, R. (2011). Social q a. *JASIST*, 62(12):2301–2312.

Groot, A. D. (1965). *Thought and choice in chess*. Mouton, The Hague.

Guo, L., Tan, E., Chen, S., Zhang, X., and Zhao, Y. E. (2009). Analyzing patterns of user content generation in online social networks. In *KDD*, pages 369–378.

Henry F, K. (1960). The application of electronic computers to factor analysis,. *Educational and Psychological Measurement*, 20:141–154.

Huotari, K. and Hamari, J. (2012). Defining gamification: A service marketing perspective. In *Proceeding of the 16th International Academic MindTrek Conference*, MindTrek '12, pages 17–22. ACM.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

Jendoubi, S., Martin, A., Liétard, L., Ben Hadji, H., and Ben Yaghlane, B. (2017). Two evidential data based models for influence maximization in twitter. *Knowledge-Based Systems*, 121:58–70.

Kao, W.-C., Liu, D.-R., and Wang, S.-W. (2010). Expert finding in question-answering websites: a novel hybrid approach. In *Proceedings of the 2010 ACM symposium on applied computing*, pages 867–871. ACM.

Kasneci, G., Gael, J. V., Stern, D. H., and Graepel, T. (2011). Cobayes: bayesian knowledge corroboration with assessors of unknown areas of expertise. In *WSDM*, pages 465–474. ACM.

Khaleghi, B., Khamis, A. M., Karray, F., and Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28–44.

King, B. (1967). Step-wise clustering procedures. *Journal of the American Statistical Association*, 62(317):86–101.

Lebart, L., Morineau, A., and Fénelon, J. (1980). *Traitement Des Données Statistiques*. Dunod.

Liu, J., Song, Y.-I., and Lin, C.-Y. (2011). Competition-based user expertise score estimation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 425–434. ACM.

Ma, Z. and Leijon, A. (2011). Bayesian estimation of beta mixture models with variational inference. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(11):2160–2173.

Ma, Z., Sun, A., Yuan, Q., and Cong, G. (2015). A tri-role topic model for domain-specific question answering. In *AAAI*, pages 224–230.

Maimon, O. and Rokach, L. (2005). *The Data Mining and Knowledge Discovery Handbook*. Springer.

Martin, A. (2009). Reliability and combination rule in the theory of belief functions. In *Information FUSION*, pages 529–536. IEEE.

Martin, A. and Osswald, C. (2007). Toward a combination rule to deal with partial conflict and specificity in belief functions theory. In *The 10th International Conference on Information FUSION*, pages 1–8. IEEE.

Martin, A. and Osswald, C. (2008). Human expert fusion for image classification. *Information & Security: An International Journal, Special issue on Fusing Uncertain, Imprecise and Conflicting Information*.

Matei, S. A., Jabal, A. A., and Bertino, E. (2017). Do sticky elites produce online knowledge of higher quality?

Morris, M. R., Teevan, J., and Panovich, K. (2010). What do people ask their social networks, and why?: a survey study of status message q&a behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1739–1748. ACM.

Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P., and Faloutsos, C. (2013). Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 886–893, New York, NY, USA. ACM.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *Computer Journal*, 26(4):354–359.

Nam, K. K., Ackerman, M. S., and Adamic, L. A. (2009). Questions in knowledge in a study of naver's question answering community. In *CHI*, pages 779–788. ACM.

Nguyen, V.-D. and Huynh, V.-N. (2016). Integrating with social network to enhance recommender system based-on dempster-shafer theory. In *CSoNet*, volume 9795 of *Lecture Notes in Computer Science*, pages 170–181. Springer.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University.

Pal, A., Chang, S., and Konstan, J. A. (2012a). Evolution of experts in question answering communities. In Breslin, J. G., Ellison, N. B., Shanahan, J. G., and Tufekci, Z., editors, *ICWSM*. The AAAI Press.

Pal, A. and Counts, S. (2011). Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 45–54. ACM.

Pal, A., Farzan, R., Konstan, J. A., and Kraut, R. E. (2011). Early detection of potential experts in question answering communities. In Konstan, J. A., Conejo, R., Marzo, J. L., and Oliver, N., editors, *UMAP*, volume 6787 of *Lecture Notes in Computer Science*, pages 231–242. Springer.

Pal, A., Harper, F. M., and Konstan, J. A. (2012b). Exploring question selection bias to identify experts and potential experts in community question answering. *ACM Trans. Inf. Syst.*, 30(2):10.

Pearson, K. (1901). On lines and planes of closest fit to points in space. *Philos. Mag*, 2,:559–572.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Provost, F. and Kohavi, R. (1998). Guest editors' introduction: On applied research in machine learning. *Machine learning*, 30(2):127–132.

Ramage, D., Hall, D. L. W., Nallapati, R., and Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, pages 248–256. ACL.

Reyni, A. (1962). *Probability Theory*. North-Holland.

Rousseeuw, P. (1987). Silhouette: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Sahu, T. P., Nagwani, N. K., and Verma, S. (2016). Multivariate beta mixture model for automatic identification of topical authoritative users in community question answering sites. *IEEE Access*, 4:5343–5355.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton.

Shah, C., Kitzie, V., and Choi, E. (2014). Modalities, motivations, and materials - investigating traditional and social online q&a services. *J. Information Science*, 40(5):669–687.

Smets, P. (1990). The combination of evidence in the transferable belief model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(5):447–458.

Smets, P. (1993). Belief functions: The disjunctive rule of combination and the generalized bayesian theorem. *Int. J. Approx. Reasoning*, 9(1):1–35.

Smets, P. (1995). The canonical decomposition of a weighted belief. In *IJCAI*, pages 1896–1901. Morgan Kaufmann.

Smets, P. (1996). Imperfect information: Imprecision-uncertainty. *Uncertainty management in information systems. from needs to solutions*, pages 225–254.

Smets, P. (2005). Decision making in the tbm: the necessity of the pignistic transformation. *Int. J. Approx. Reasoning*, 38(2):133–147.

Smets, P. and Kennes, R. (1994). The transferable belief model. *Artificial Intelligence*, 66:191–234.

Song, S., Tian, Y., Han, W., Que, X., and Wang, W. (2013). Leading users detecting model in professional community question answering services. In *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing*, pages 1302–1307. IEEE.

Srba, I. and Bielikova, M. (2016a). A comprehensive survey and classification of approaches for community question answering. *TWEB*, 10(3):18:1–18:63.

Srba, I. and Bielikova, M. (2016b). Why is stack overflow failing? preserving sustainability in community question answering. *IEEE Software*, 33(4):80–89.

Tang, X. and Yang, C. C. (2012). Ranking user influence in healthcare social media. *ACM TIST*, 3(4):73.

Todeschini, R. (1997). Data correlation, number of significant principal components and shape of molecules. the k correlation index. *Anal. Chim. Acta*, 348,:419–430.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison Wesley, Massachusetts.

Tuna, T., Akbas, E., Aksoy, A., Canbaz, M. A., Karabiyik, U., Gonen, B., and Aygun, R. (2016). User characterization for online social networks. *Social Netw. Analys. Mining*, 6(1):104:1–104:28.

Van Dijk, D., Tsagkias, M., and de Rijke, M. (2015). Early detection of topical expertise in community question answering. In Baeza-Yates, R. A., Lalmas, M., Moffat, A., and Ribeiro-Neto, B. A., editors, *SIGIR*, pages 995–998. ACM.

Voss, J. F. and Wiley, J. (2006). Expertise in history. In *the cambridge handbook of expertise and expert performance*, pages 569–584.

Webster, N. (1968). *The World Publishing Company*. Cleveland, OH, Springfield.

White, A. J., Chan, J., Hayes, C., and Murphy, T. B. (2012). Mixed membership models for exploring user roles in online fora. In *ICWSM*.

Yang, J., Tao, K., Bozzon, A., and Houben, G.-J. (2014). Sparrows and owls: Characterisation of expert behaviour in stackoverflow. In Dimitrova, V., Kuflik, T., Chin, D., Ricci, F., Dolog, P., and Houben, G.-J., editors, *UMAP*, volume 8538 of *Lecture Notes in Computer Science*, pages 266–277. Springer.

Yang, J. and Wei, X. (2009). Seeking and offering expertise across categories: A sustainable mechanism works for baidu knows. In Adar, E., Hurst, M., Finin, T., Glance, N. S., Nicolov, N., and Tseng, B. L., editors, *ICWSM*. The AAAI Press.

Yang, L., Qiu, M., Gottipati, S., Zhu, F., Jiang, J., Sun, H., and Chen, Z. (2013). Cqarank: jointly model topics and expertise in community question answering. In He, Q., Iyengar, A., Nejdl, W., Pei, J., and Rastogi, R., editors, *CIKM*, pages 99–108. ACM.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8:338–353.

Zhang, J., Ackerman, M. S., and Adamic, L. (2007). Expertise networks in online communities: structure and algorithms. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 221–230. ACM Press.

Zhou, K., Martin, A., Pan, Q., and Liu, Z. (2016). Ecmdd: Evidential c-medoids clustering with multiple prototypes. *Pattern Recognition*, 60:239–257.

Bibliography

Ackerman, M. and McDonald, D. (1996). AnswerGarden 2: Merging organizational memory with collaborative help. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work*, pages 97–105.

Adamic, L. A., Zhang, J., Bakshy, E., and Ackerman, M. S. (2008). Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, pages 665–674. ACM.

Aslay, C., O'Hare, N., Aiello, L. M., and Jaimes, A. (2013). Competition-based networks for expert finding. In Jones, G. J. F., Sheridan, P., Kelly, D., de Rijke, M., and Sakai, T., editors, *SIGIR*, pages 1033–1036. ACM.

Attiaoui, D., Martin, A., and Ben Yaghlane, B. (2015). Détection des experts dans un cadre incertain. In *Plate-forme Intelligence Artificielle*.

Attiaoui, D., Martin, A., and Ben Yaghlane, B. (2017a). Belief measure of expertise for experts detection in question answering communities: case study stack overflow. In *21st International Conference on Knowledge-Based and Intelligent Information Engineering Systems-KES*.

Attiaoui, D., Martin, A., and Ben Yaghlane, B. (2017b). Belief temporal analysis of expert users: case study stack overflow. In *19th International Conference on Big Data Analytics and Knowledge Discovery - DaWaK*.

Balog, K. and de Rijke, M. (2009). Combining candidate and document models for expert search. In Voorhees, E. M. and Buckland, L. P., editors, *TREC*, volume Special Publication 500-277. National Institute of Standards and Technology (NIST).

Bouguessa, M., Dumoulin, B., and Wang, S. (2008). Identifying authoritative actors in question-answering forums: the case of yahoo! answers. In Li, Y., Liu, B., and Sarawagi, S., editors, *KDD*, pages 866–874. ACM.

Bouguessa, M. and Romdhane, L. B. (2015). Identifying authorities in online communities. *ACM TIST*, 6(3):30.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics Simulation and Computation*, 3(1):1–27.

Chan, J., Hayes, C., and Daly, E. (2010). Decomposing discussion forums using common user roles. *Proceedings of the WebSci10: Extending the Frontiers of Society OnLine*.

Chase, W., Simon, H. A., and Chase, W. (1973). *The mind's eye in chess*. Academic Press, New York.

Chase, W. G. and Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4:55–81.

Chebbah, M., Martin, A., and Ben Yaghlane, B. (2015). Combining partially independent belief functions. *Decision Support Systems*, 73:37–46.

Chen, Y., Ho, T.-H., and Kim, Y.-M. (2010). Knowledge market design: A field experiment at google answers. *Journal of Public Economic Theory*, 12(4):641–664.

Chirag Shah, S. O. and Oh, J. S. (2009). Research agenda for social q a. volume 31 of *library and information science research journal*, page 205âĂŞ209. Springer.

Choi, E., Kitzie, V., and Shah, C. (2012). Developing a typology of online q&a models and recommending the right model for each question type. *Proceedings of the Association for Information Science and Technology*, 49(1):1–4.

Cleuziou, G. (2004). *Une méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information. (A Clustering method for rules learning and information retrieval).* PhD thesis, University of Orléans, France.

Cumming, G., Fidler, F., and Vaux, D. L. (2007). Error bars in experimental biology. *The Journal of cell biology*, 177(1):7–11.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1(2):224–227.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. In *Classic Works of the Dempster-Shafer Theory of Belief Functions*, volume 219.

Denoeux, T. (1995). A k-nearest neighbor classification rule based on dempster-shafer theory. *IEEE Trans. Systems, Man, and Cybernetics*, 25(5):804–813.

Dlala, I. O., Attiaoui, D., Martin, A., and Ben Yaghlane, B. (2015). Trolls identification within an uncertain framework. abs/1501.05272.

Dubois, D. and Prade, H. (1988). Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence*, 4:244–264.

Dubois, D. and Prade, H. (2015). Possibility theory and its applications: Where do we stand? In *Handbook of Computational Intelligence*, pages 31–60. Springer.

Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4:95–104.

Ericsson, K. A. (2006). The cambridge handbook of expertise and expert performance.

Essaid, A., Martin, A., Smits, G., and Ben Yaghlane, B. (2014). A distance-based decision in the credal level. *International Conference on Artificial Intelligence and Symbolic Computation AISC*.

Estivill-Castro, V. and Yang, J. (2000). Fast and robust general purpose clustering algorithms. In *PRICAI*, pages 208–218.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). Advances in knowledge discovery and data mining.

Forestier, M., Stavrianou, A., Velcin, J., and Zighed, D. A. (2012). Roles in social networks: Methodologies and research issues. *Web Intelligence and Agent Systems*, 10(1):117–133.

Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588.

Furlan, B., Nikolic, B., and Milutinovic, V. (2013). A survey and evaluation of state-of-the-art intelligent question routing systems. *Int. J. Intell. Syst.*, 28(7):686–708.

Furtado, A., Andrade, N., Oliveira, N., and Brasileiro, F. V. (2013). Contributor profiles, their dynamics, and their importance in five q&a sites. pages 1237–1252. ACM.

Gazan, R. (2011). Social q a. *JASIST*, 62(12):2301–2312.

Groot, A. D. (1965). *Thought and choice in chess*. Mouton, The Hague.

Guo, L., Tan, E., Chen, S., Zhang, X., and Zhao, Y. E. (2009). Analyzing patterns of user content generation in online social networks. In *KDD*, pages 369–378.

Henry F, K. (1960). The application of electronic computers to factor analysis,. *Educational and Psychological Measurement*, 20:141–154.

Huotari, K. and Hamari, J. (2012). Defining gamification: A service marketing perspective. In *Proceeding of the 16th International Academic MindTrek Conference*, MindTrek '12, pages 17–22. ACM.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

Jendoubi, S., Martin, A., Liétard, L., Ben Hadji, H., and Ben Yaghlane, B. (2017). Two evidential data based models for influence maximization in twitter. *Knowledge-Based Systems*, 121:58–70.

Kao, W.-C., Liu, D.-R., and Wang, S.-W. (2010). Expert finding in question-answering websites: a novel hybrid approach. In *Proceedings of the 2010 ACM symposium on applied computing*, pages 867–871. ACM.

Kasneci, G., Gael, J. V., Stern, D. H., and Graepel, T. (2011). Cobayes: bayesian knowledge corroboration with assessors of unknown areas of expertise. In *WSDM*, pages 465–474. ACM.

Khaleghi, B., Khamis, A. M., Karray, F., and Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28–44.

King, B. (1967). Step-wise clustering procedures. *Journal of the American Statistical Association*, 62(317):86–101.

Lebart, L., Morineau, A., and Fénelon, J. (1980). *Traitement Des Données Statistiques*. Dunod.

Liu, J., Song, Y.-I., and Lin, C.-Y. (2011). Competition-based user expertise score estimation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 425–434. ACM.

Ma, Z. and Leijon, A. (2011). Bayesian estimation of beta mixture models with variational inference. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(11):2160–2173.

Ma, Z., Sun, A., Yuan, Q., and Cong, G. (2015). A tri-role topic model for domain-specific question answering. In *AAAI*, pages 224–230.

Maimon, O. and Rokach, L. (2005). *The Data Mining and Knowledge Discovery Handbook*. Springer.

Martin, A. (2009). Reliability and combination rule in the theory of belief functions. In *Information FUSION*, pages 529–536. IEEE.

Martin, A. and Osswald, C. (2007). Toward a combination rule to deal with partial conflict and specificity in belief functions theory. In *The 10th International Conference on Information FUSION*, pages 1–8. IEEE.

Martin, A. and Osswald, C. (2008). Human expert fusion for image classification. *Information & Security: An International Journal, Special issue on Fusing Uncertain, Imprecise and Conflicting Information*.

Matei, S. A., Jabal, A. A., and Bertino, E. (2017). Do sticky elites produce online knowledge of higher quality?

Morris, M. R., Teevan, J., and Panovich, K. (2010). What do people ask their social networks, and why?: a survey study of status message q&a behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1739–1748. ACM.

Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P., and Faloutsos, C. (2013). Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 886–893, New York, NY, USA. ACM.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *Computer Journal*, 26(4):354–359.

Nam, K. K., Ackerman, M. S., and Adamic, L. A. (2009). Questions in knowledge in a study of naver's question answering community. In *CHI*, pages 779–788. ACM.

Nguyen, V.-D. and Huynh, V.-N. (2016). Integrating with social network to enhance recommender system based-on dempster-shafer theory. In *CSoNet*, volume 9795 of *Lecture Notes in Computer Science*, pages 170–181. Springer.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University.

Pal, A., Chang, S., and Konstan, J. A. (2012a). Evolution of experts in question answering communities. In Breslin, J. G., Ellison, N. B., Shanahan, J. G., and Tufekci, Z., editors, *ICWSM*. The AAAI Press.

Pal, A. and Counts, S. (2011). Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 45–54. ACM.

Pal, A., Farzan, R., Konstan, J. A., and Kraut, R. E. (2011). Early detection of potential experts in question answering communities. In Konstan, J. A., Conejo, R., Marzo, J. L., and Oliver, N., editors, *UMAP*, volume 6787 of *Lecture Notes in Computer Science*, pages 231–242. Springer.

Pal, A., Harper, F. M., and Konstan, J. A. (2012b). Exploring question selection bias to identify experts and potential experts in community question answering. *ACM Trans. Inf. Syst.*, 30(2):10.

Pearson, K. (1901). On lines and planes of closest fit to points in space. *Philos. Mag*, 2,:559–572.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Provost, F. and Kohavi, R. (1998). Guest editors' introduction: On applied research in machine learning. *Machine learning*, 30(2):127–132.

Ramage, D., Hall, D. L. W., Nallapati, R., and Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, pages 248–256. ACL.

Reyni, A. (1962). *Probability Theory*. North-Holland.

Rousseeuw, P. (1987). Silhouette: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Sahu, T. P., Nagwani, N. K., and Verma, S. (2016). Multivariate beta mixture model for automatic identification of topical authoritative users in community question answering sites. *IEEE Access*, 4:5343–5355.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton.

Shah, C., Kitzie, V., and Choi, E. (2014). Modalities, motivations, and materials - investigating traditional and social online q&a services. *J. Information Science*, 40(5):669–687.

Smets, P. (1990). The combination of evidence in the transferable belief model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(5):447–458.

Smets, P. (1993). Belief functions: The disjunctive rule of combination and the generalized bayesian theorem. *Int. J. Approx. Reasoning*, 9(1):1–35.

Smets, P. (1995). The canonical decomposition of a weighted belief. In *IJCAI*, pages 1896–1901. Morgan Kaufmann.

Smets, P. (1996). Imperfect information: Imprecision-uncertainty. *Uncertainty management in information systems. from needs to solutions*, pages 225–254.

Smets, P. (2005). Decision making in the tbm: the necessity of the pignistic transformation. *Int. J. Approx. Reasoning*, 38(2):133–147.

Smets, P. and Kennes, R. (1994). The transferable belief model. *Artificial Intelligence*, 66:191–234.

Song, S., Tian, Y., Han, W., Que, X., and Wang, W. (2013). Leading users detecting model in professional community question answering services. In *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing*, pages 1302–1307. IEEE.

Srba, I. and Bielikova, M. (2016a). A comprehensive survey and classification of approaches for community question answering. *TWEB*, 10(3):18:1–18:63.

Srba, I. and Bielikova, M. (2016b). Why is stack overflow failing? preserving sustainability in community question answering. *IEEE Software*, 33(4):80–89.

Tang, X. and Yang, C. C. (2012). Ranking user influence in healthcare social media. *ACM TIST*, 3(4):73.

Todeschini, R. (1997). Data correlation, number of significant principal components and shape of molecules. the k correlation index. *Anal. Chim. Acta*, 348,:419–430.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison Wesley, Massachusetts.

Tuna, T., Akbas, E., Aksoy, A., Canbaz, M. A., Karabiyik, U., Gonen, B., and Aygun, R. (2016). User characterization for online social networks. *Social Netw. Analys. Mining*, 6(1):104:1–104:28.

Van Dijk, D., Tsagkias, M., and de Rijke, M. (2015). Early detection of topical expertise in community question answering. In Baeza-Yates, R. A., Lalmas, M., Moffat, A., and Ribeiro-Neto, B. A., editors, *SIGIR*, pages 995–998. ACM.

Voss, J. F. and Wiley, J. (2006). Expertise in history. In *the cambridge handbook of expertise and expert performance*, pages 569–584.

Webster, N. (1968). *The World Publishing Company*. Cleveland, OH, Springfield.

White, A. J., Chan, J., Hayes, C., and Murphy, T. B. (2012). Mixed membership models for exploring user roles in online fora. In *ICWSM*.

Yang, J., Tao, K., Bozzon, A., and Houben, G.-J. (2014). Sparrows and owls: Characterisation of expert behaviour in stackoverflow. In Dimitrova, V., Kuflik, T., Chin, D., Ricci, F., Dolog, P., and Houben, G.-J., editors, *UMAP*, volume 8538 of *Lecture Notes in Computer Science*, pages 266–277. Springer.

Yang, J. and Wei, X. (2009). Seeking and offering expertise across categories: A sustainable mechanism works for baidu knows. In Adar, E., Hurst, M., Finin, T., Glance, N. S., Nicolov, N., and Tseng, B. L., editors, *ICWSM*. The AAAI Press.

Yang, L., Qiu, M., Gottipati, S., Zhu, F., Jiang, J., Sun, H., and Chen, Z. (2013). Cqarank: jointly model topics and expertise in community question answering. In He, Q., Iyengar, A., Nejdl, W., Pei, J., and Rastogi, R., editors, *CIKM*, pages 99–108. ACM.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8:338–353.

Zhang, J., Ackerman, M. S., and Adamic, L. (2007). Expertise networks in online communities: structure and algorithms. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 221–230. ACM Press.

Zhou, K., Martin, A., Pan, Q., and Liu, Z. (2016). Ecmdd: Evidential c-medoids clustering with multiple prototypes. *Pattern Recognition*, 60:239–257.