



PARSING



MATCHING



EQUALITY



SEARCH

## LIVRE BLANC

Décembre 2014



### Le marché de l'emploi vu à travers les CV

Une application de text-analytics réalisée avec LEA-CV

## Introduction

L'analyse des tendances du marché de l'emploi correspond à l'évidence à une nécessité, surtout en période de tension comme depuis plusieurs années en France et en Europe. De très nombreuses études, réalisées par autant d'organismes, abordent ce sujet. Une approche très souvent utilisée est de partir par l'analyse des offres, ce qui donne une image des besoins des entreprises. Plus rare, du moins à notre connaissance, est une approche partant de l'analyse des CV, qui reflète les besoins non pas des entreprises, mais bien des candidats à l'emploi.

Au moins deux difficultés doivent être abordées si l'on veut suivre une telle approche partant des CV : d'une part, en recueillir un grand nombre pour assurer une bonne représentativité, d'autre part, et surtout, être capable d'analyser ces CV qui sont des textes hétérogènes, pour en extraire et normaliser les données nécessaires à l'étude. Sur le premier point, et dans la mouvance du Big Data, de nombreuses méthodes se sont développées pour analyser des masses importantes de données. Sur le deuxième point, les méthodes déjà bien maîtrisées depuis plusieurs années du « data analytics » se sont aujourd'hui enrichies du « text analytics », qui permettent une approche statistique et synthétique des informations non structurées « cachées » dans les textes.

« Analytics » : depuis quelques années, le mot est à la mode : une simple recherche sur Google Trends, qui mesure la fréquence des questions posées au moteur de recherche, montre une forte croissance de l'intérêt autour de cette notion depuis 7 ou 8 ans :

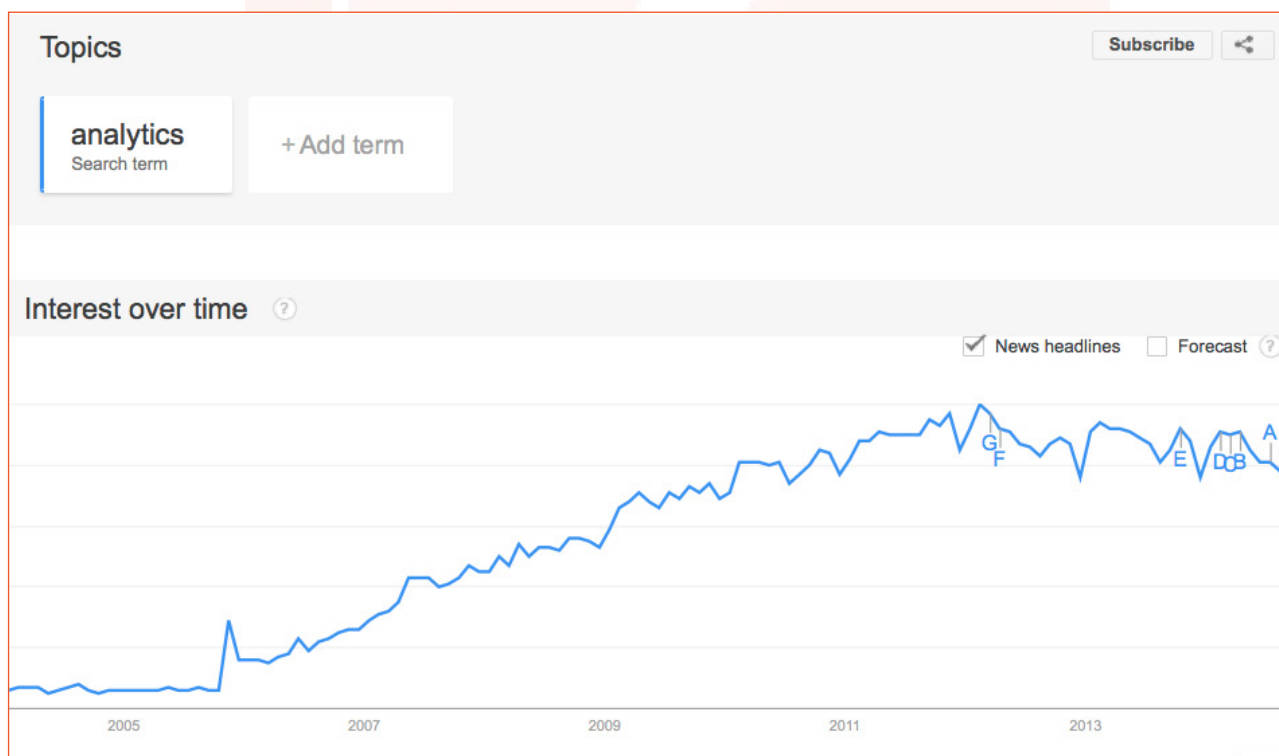


Figure 1 - Importance du mot « analytics » dans Google Trends

Cette tendance est liée notamment à l'explosion du « Big Data », qui rend possible toute une série de nouvelles applications basées sur l'analyse des grandes masses de données. Ce sont généralement des données structurées, provenant de toutes sortes de sources « classiques » internes ou externes à l'entreprise. Avec l'explosion des objets connectés, ce sont pratiquement tous les secteurs d'activités qui sont concernés : la banque et l'assurance, la santé, les transports, les gouvernements, etc. Dans l'organisation de l'entreprise, ce sont, pour le moment, surtout les directions des Etudes marketing, de l'Intelligence économique, de la relation clients qui sont les principaux utilisateurs de ces technologies.

On en parle moins dans le monde des Ressources Humaines et de l'emploi, sans doute parce que les données y sont très largement de nature textuelle, plus que dans d'autres activités de l'entreprise. L'analyse de données doit alors faire majoritairement appel à ces technologies du text-mining, qui est le pendant dans le monde des données non structurées, de ce qu'est le data-mining dans le monde des données structurées.

Ces technologies sont en train d'arriver à maturité et suscitent de plus en plus d'intérêt, comme le montre la recherche sur Google Trends ci-dessous.

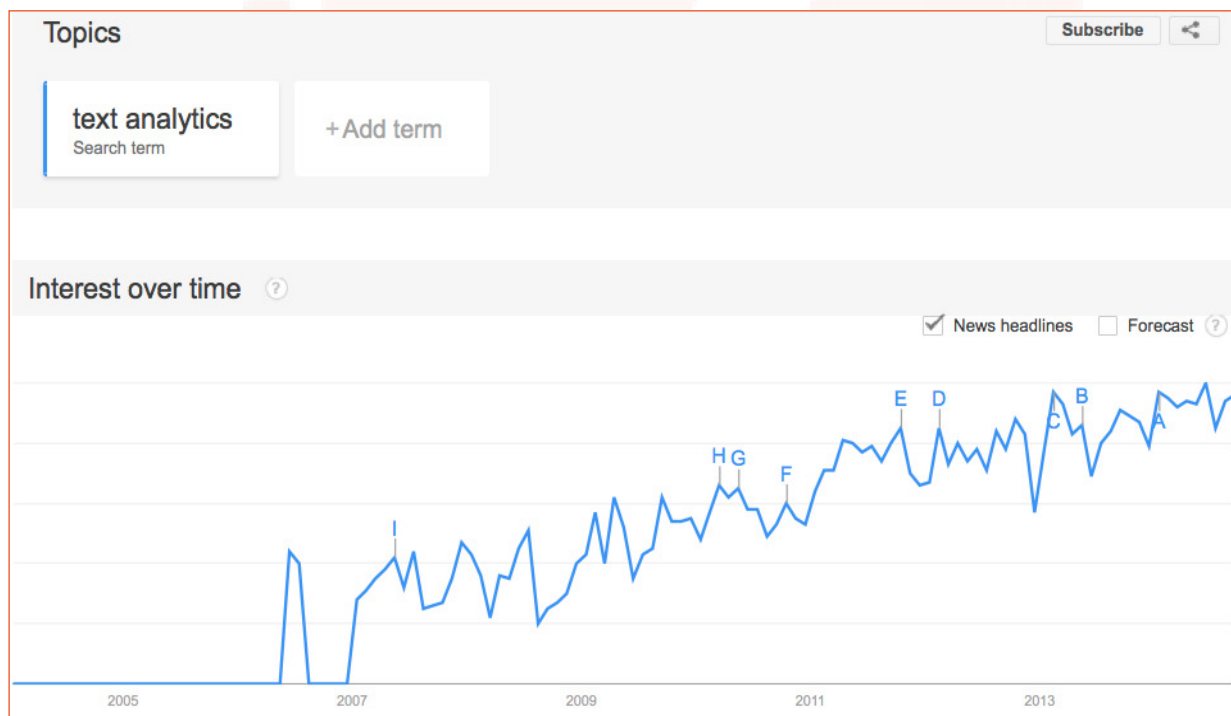


Figure 2 : Importance croissante de « text analytics »

On présente, ici, un exemple d'application basée sur ce « text analytics » dans le monde des Ressources Humaines. L'idée est d'exploiter une base de données de CV pour voir si on peut en tirer des indications sur le marché de l'emploi, le parcours professionnel des demandeurs, les adéquations entre formations et emplois obtenus, etc.


## Présentation globale de la méthode


Cette étude est basée sur l'analyse d'un corpus de 40 000 Curriculum Vitae récents, recueillis sur une région en France, tous métiers confondus. On ne se prononce pas ici sur la représentativité de ce corpus à l'échelle nationale, mais on peut sans doute considérer qu'il représente bien une synthèse de la situation de ce bassin d'emploi particulier. Pour des raisons de confidentialité des données tous les résultats présentés ici sont anonymisés, aucune référence à un CV individuel, encore moins à un nom ou identification d'une personne n'étant donnée. On ne donnera pas non plus d'indication sur les villes ou régions concernées.


Il s'agit plutôt de montrer comment des outils d'analyse de textes, appliqués à un corpus de CV, permettent de déboucher sur un véritable outil de text-analytics. Les analyses qui suivent ont été réalisées avec le module « analytics » de LeaCV.


### Les informations extraites


Pour cette application, les informations suivantes ont été extraites des textes de CV par l'analyseur linguistique :

 **Le poste visé** : c'est généralement le poste ou la fonction indiqué en tête du CV, ou dans son titre. Il arrive parfois que certains CV, ne comportent pas réellement d'en-tête ou de titre. On prend alors l'intitulé du dernier poste mentionné dans le corps du CV.

 **Les dernier et avant dernier poste** : ces deux informations sont extraites du corps du CV. Cela suppose d'avoir préalablement identifié la structure chronologique de la carrière du candidat, ce qui n'est pas trivial, toutes sortes de variantes dans l'expression des dates et périodes d'emploi pouvant être rencontrées dans les CV : ordre chronologique croissant ou décroissant, formats de dates variés, « trous » dans les périodes d'emploi, etc.

 **Le niveau d'expérience** : c'est une métadonnée calculée à partir de l'analyse de la chronologie, en fonction du nombre d'années d'expérience.

 **Les compétences** : les compétences intéressantes dépendent à l'évidence des secteurs d'activité. Dans cette expérimentation on s'est contenté d'identifier des compétences très génériques, comme la connaissance d'outils bureautiques et informatiques.

 **Les diplômes** : reconnaître un diplôme nécessite d'identifier plusieurs éléments : son intitulé (CAP, BAC, BTS, Licence, Master, Doctorat, Ingénieur, etc.), l'identification de l'établissement ayant délivré le diplôme souvent indispensable, comme par exemple pour les écoles d'ingénieurs, ou encore l'année d'obtention du diplôme qui est un indicateur intéressant. Mais, c'est la discipline du diplôme qui est la plus variable (une licence en chimie moléculaire, un BTS en informatique de gestion...). Il n'est, en pratique, pas possible de se baser sur une liste fermée de diplômes, ne serait ce que parce qu'il s'en crée de nouveaux tous les jours. Certaines de ces méta-données sont très variables, au moins dans leur formulation par les candidats qui ne reprennent pas forcément la désignation « officielle » de leur diplôme, comme le montre la liste de BTS ci-après :

**Liste des BTS**

BTS Action Commerciale	BTS Comptabilité et Gestion des Entreprises
BTS Assistant de Manager	BTS Comptabilité et Gestion des Organisations
BTS Assistante de Direction	BTS Comptabilité Gestion
BTS Assistante de Gestion	BTS Comptabilité Gestion des Organisations
BTS Assistante de Gestion PME-PMI	BTS Comptabilité-Gestion
BTS Assistante de Manager	BTS Conception de Produits Industriels
BTS Assistante Secrétaire Trilingue	BTS Electronique
BTS Assurance	BTS Electrotechnique
BTS CGO (Comptabilité et Gestion des Organisations)	BTS Force de Vente
BTS Commerce International	BTS Informatique de Gestion
BTS Communication	BTS Maintenance Industrielle
BTS Communication des Entreprises	BTS Management des Unités Commerciales (MUC)
BTS Comptabilité	BTS Mécanique et Automatismes Industriels
BTS Comptabilité et Gestion	BTS Négociation et Relation Client

**Le meilleur diplôme** : c'est celui reconnu comme ayant le niveau le plus élevé.

**Les entreprises** : la reconnaissance des noms d'entreprises, souvent considérée comme un problème classique, voire résolu, de l'extraction d'entités nommées est en fait, si l'on veut une bonne qualité de résultats, une tâche difficile. D'une part, il est impossible de se baser sur une liste fermée de noms d'entreprises. Même si de nombreux référentiels existent, notamment dans le monde de l'open data, ils sont toujours incomplets, ne serait-ce que parce qu'il se crée de nouvelles entreprises en permanence. Un autre type de difficulté vient de l'ambiguïté très fréquente entre nom d'entreprise et nom de produit. «Coca-Cola», peut, selon le contexte, désigner la boisson ou la société. C'est le contexte qui permet de lever l'ambiguïté, par exemple dans un CV « 2010 - 2012 : Samsung ...» désignera l'entreprise alors que dans « 2010 - 2012 j'ai travaillé sur l'ergonomie du Samsung Galaxy S5 » c'est le produit qui est désigné. Par ailleurs, les noms d'entreprises sont souvent ambigus avec des noms de personnes (centre Edouard Leclerc), des mots du vocabulaire commun (Axa, Orange, etc.).

**Les écoles de même que pour les entreprises** : aucun référentiel ne pouvant être en permanence à jour, avec notamment les entreprises à l'étranger, celle n'existant plus, etc., c'est l'analyse du contexte qui permet d'identifier les noms des écoles.

## La gestion des variantes

Une des difficultés principale est la gestion des variantes. Lors de l'extraction des métadonnées, il est important de regrouper différentes formes synonymes, faute de quoi les traitements statistiques ultérieurs seraient faussés. Par exemple :

**Les césures de mots concaténés** : la façon d'orthographier des mots plus ou moins techniques, ou des noms de produits est très variable. En particulier, les mots construits comme des mots composés concaténés sont écrits de bien des façons. Voici par exemple différentes façons d'écrire « powerpoint » relevées dans des CV :

Forme : Power Point, Powerpoint, POWER POINT, PowerPoint, Power point, power point, powerpoint, POWERPOINT, Microsoft PowerPoint, Power Point, PowerPoint 2007, PowerPoINT, Microsoft Powerpoint, Powerpoint 2007, MS PowerPoint, poWerpoinT, power point, power Point

De la même manière on peut considérer que « technicien de maintenance » et « agent de maintenance », avec leur variantes avec ou sans préposition avec ou sans majuscules, sont en fait synonymes et qu'il faut les regrouper lors d'une analyse statistique.

Forme : Technicien de maintenance, technicien de maintenance, Technicien maintenance, Technicien de Maintenance, Agent de maintenance, Agent de Maintenance, TECHNICIEN DE MAINTENANCE, Technicien De Maintenance, Technicien Maintenance, agent de maintenance, Technicien en Maintenance, agent maintenance, technicien de Maintenance, AGENT DE MAINTENANCE, AGENT DE MAINTENANCE, Technicien en maintenance, technicien maintenance

Les noms des sociétés sont également très variables, même celles qui paraissent les plus connues. A titre d'exemple, on voit ci-dessous un grand nombre de variantes rencontrées dans des CV pour désigner les magasins E. Leclerc :

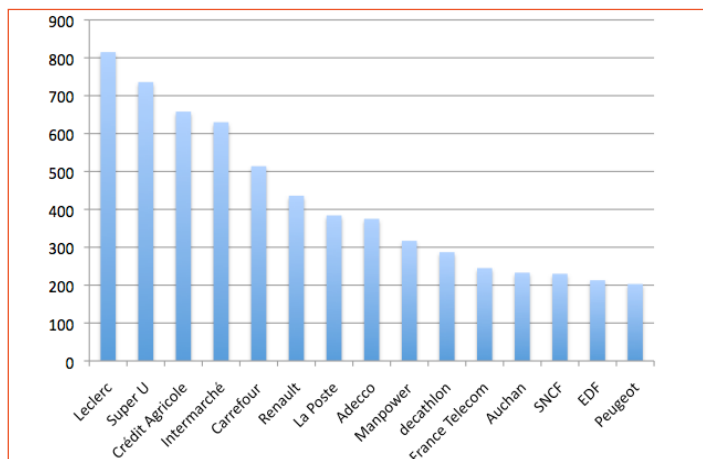
Forme : Centre Leclerc, Leclerc, LECLERC, E. LECLERC, E leclerc, Centre E. Leclerc, E-Leclerc, e-leclerc, E. Leclerc, Centre E LECLERC, CENTRE LECLERC, supermarché LECLERC, centre leclerc, CENTRE E. LECLERC, groupe Leclerc, centre Leclerc, Magasin Leclerc, E Leclerc, Centre Commercial E. Leclerc, Groupe LECLERC, Hypermarché E. Leclerc, Centre E. LECLERC, Centre Commercial Leclerc, CENTRE E LECLERC, Supermarché LECLERC, Hypermarché E. LECLERC, centre LECLERC, centre E. Leclerc, magasin Leclerc, centre E-Leclerc, Groupe Leclerc, SOCIETE LECLERC, leclerc, Centre E-Leclerc, E-LECLERC, Centre LECLERC, magasin E. LECLERC, Centre Commercial LECLERC, Centre E Leclerc, E LECLERC, Société LECLERC, Magasin E. Leclerc, hypermarché E. Leclerc, magasin LECLERC, CENTRE E-LECLERC, Magasin LECLERC, E. LECLERC, centre commercial E. LECLERC, CENTRE COMMERCIAL E. LECLERC, Société Leclerc, centre e leclerc, centre E. LECLERC, centre commercial E. Leclerc

On comprend donc l'importance de la normalisation des différentes formes rencontrées dans les textes, indispensable avant de tenter toute analyse statistique. C'est le point critique dans les système de text-analytics : de nombreuses plateformes proposent des résultats statistiques, agréablement présentés dans des dashboards spectaculaires, mais qui se basent sur des données non normalisées, extraites sur la base de mots clés sans sémantique particulière et conduisant donc souvent à des résultats sans grand intérêt, voire faux.

## Résultats

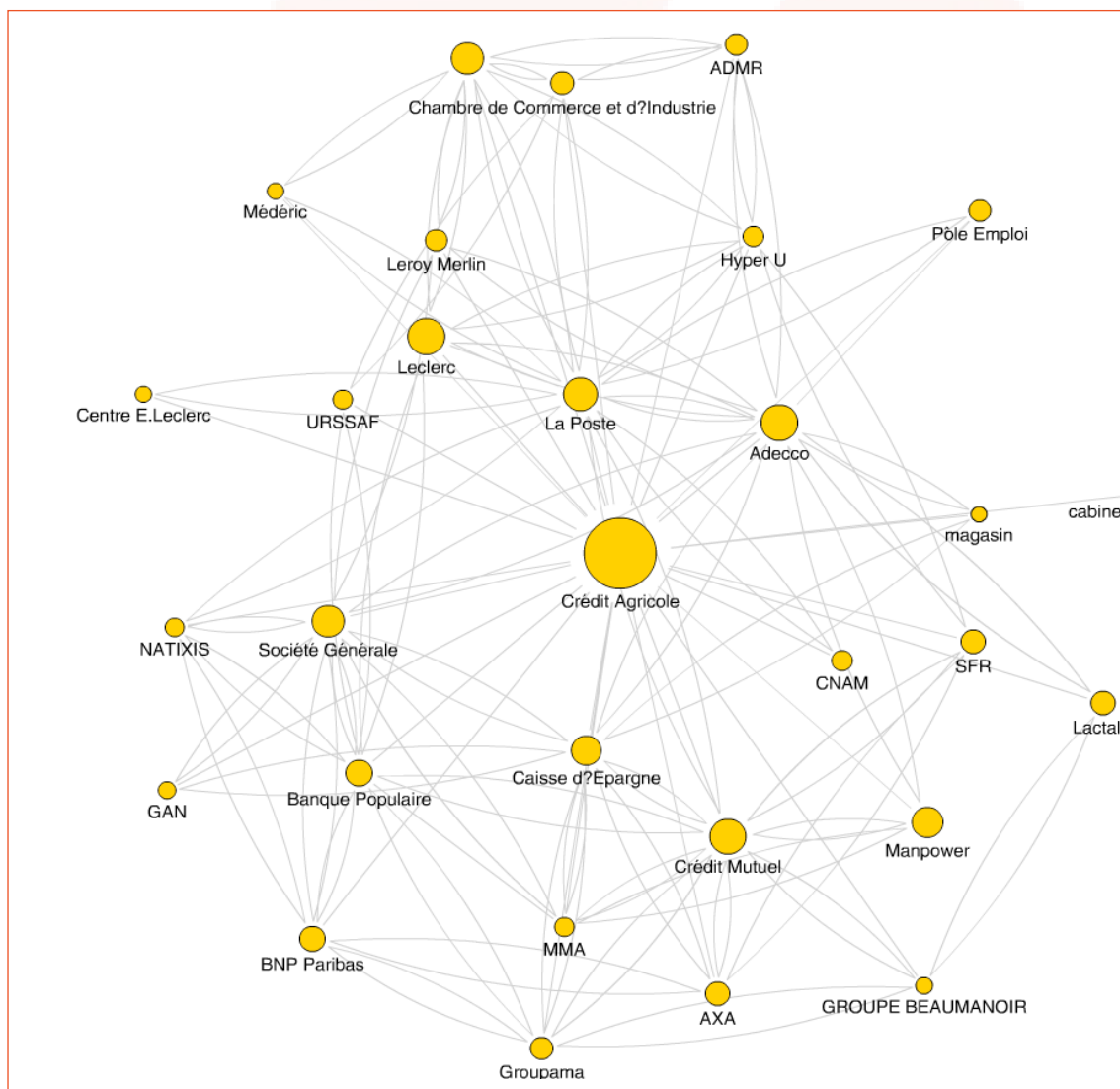
### Une vue globale des demandeurs sur le bassin d'emploi

L'exploitation de la base de données ainsi constituée permet toutes sortes d'analyses. Au niveau global des 40 000 CV, on peut, par exemple, identifier les entreprises les plus fréquemment citées :

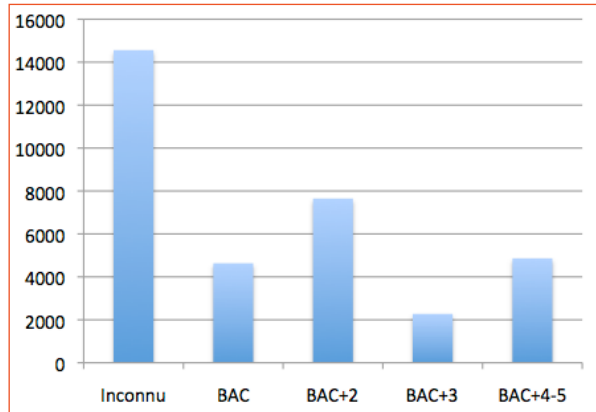


L'importance du secteur de la grande distribution est frappante. Cela ne signifie pas que ces entreprises ont, à un instant donné, le plus d'offres d'emploi, mais que ce sont celles qui sont le plus fréquemment citées dans les historiques de carrières des CV. Une analyse plus approfondie, qui n'a pas été faite ici, serait d'analyser le turn-over par société, à partir des dates de début et de fin de chaque période d'emploi. On a néanmoins une information intéressante, qui est, que sur ce bassin d'emploi, une proportion très significative des demandeurs sont passés au moins une fois, dans leur vie professionnelle, par une entreprise du secteur de la distribution.

Une analyse complémentaire est de rechercher les co-citations de noms d'entreprises dans les CV, comme montré sur le graphe ci-dessous, qui s'interprète de la façon suivante : le point central représente l'entreprise sur laquelle porte l'analyse ; la surface des cercles est proportionnelle au nombre de CV mentionnant l'entreprise, les arcs représentent une co-citation, c'est à dire qu'ils représentent une personne ayant travaillé dans les deux entreprises reliées par l'arc. On voit ainsi apparaître des affinités entre entreprises, en ce sens qu'elles ont en commun d'avoir employé, à un moment ou à un autre, les mêmes salariés.



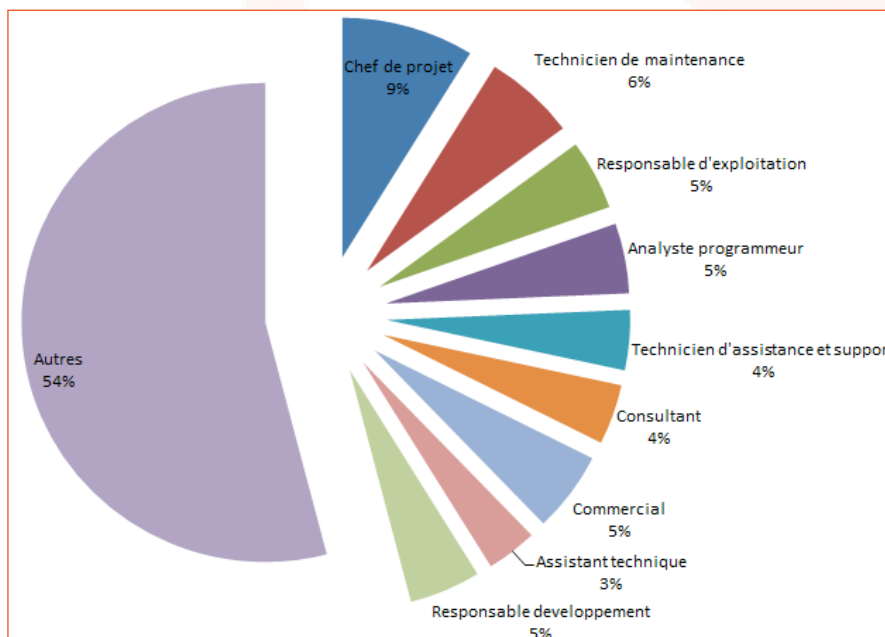
D'autres analyses globales sont faciles à faire, comme par exemple la répartition des niveaux de diplômes ci-dessous.



On note l'importance des CV dans lesquels l'analyseur n'a pas identifié le niveau du diplôme. Cela peut provenir soit du fait que le CV ne mentionne aucun diplôme, soit qu'il est effectivement mentionné, mais que l'analyseur n'a pas pu lui associer un niveau. On peut considérer, par approximation, que ce sont des diplômes peu connus, de niveau BAC ou moins.

### Des analyses sectorielles

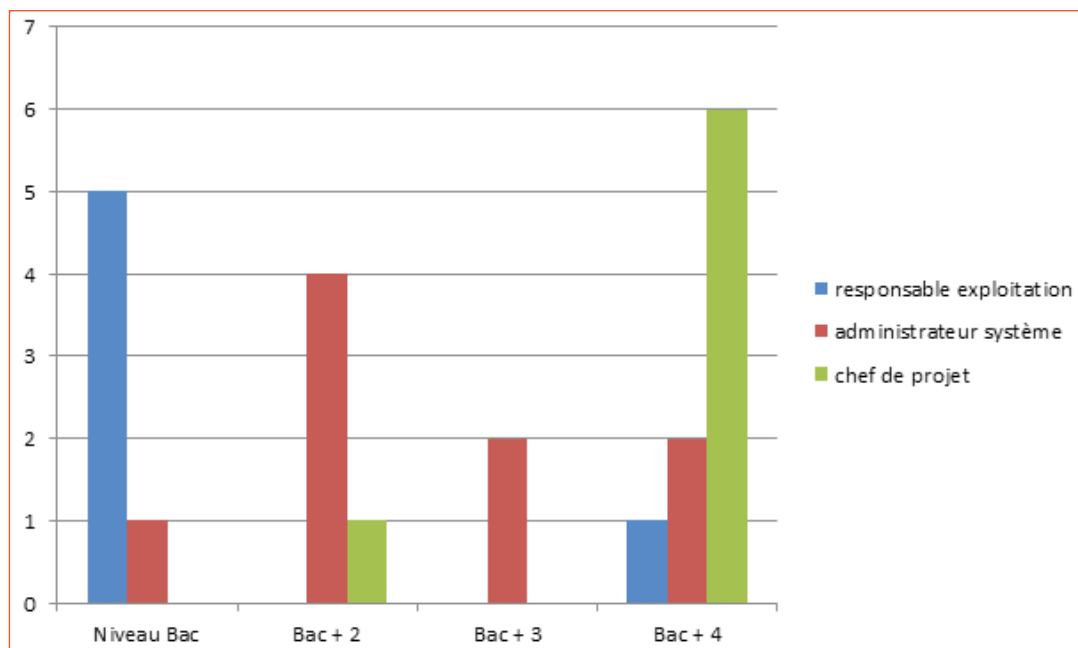
Au delà d'une analyse globale sur l'ensemble du bassin d'emploi, des analyses sectorielles peuvent être intéressantes. Par exemple, à partir d'une recherche sur environ 950 postes dans le domaine général de l'informatique, l'outil permet d'obtenir une répartition des CV en fonction des postes recherchés.



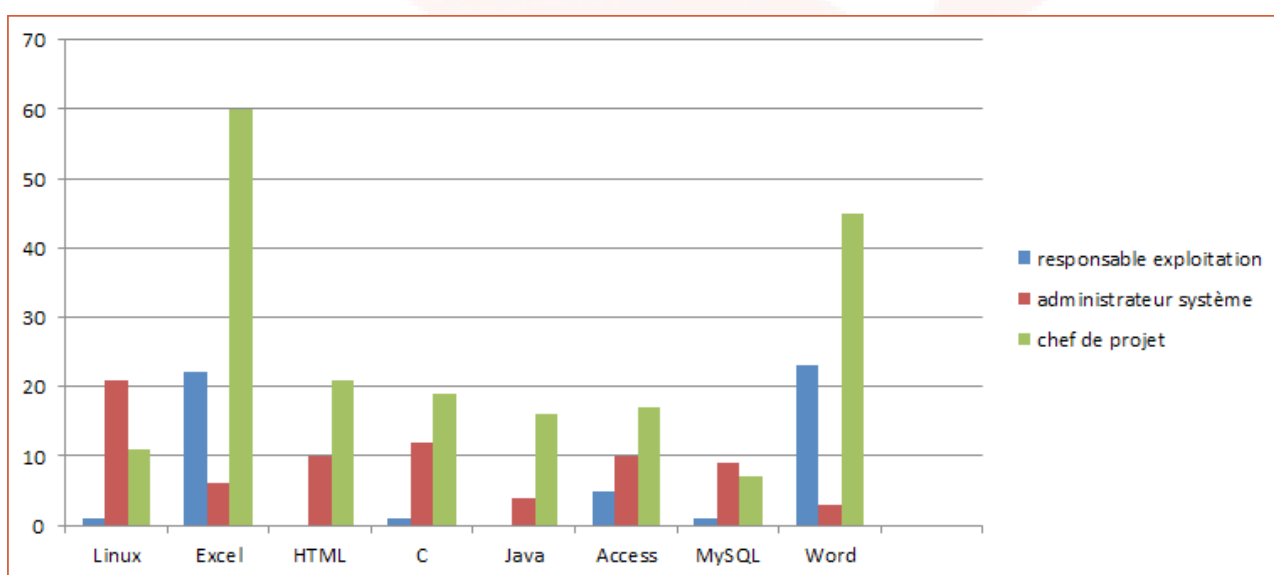
On voit que l'échantillon recueilli couvre essentiellement des postes techniques, très peu de postes de commerciaux ne sont présents. Cela ne prétend pas être représentatif de la situation à l'échelle nationale.



Diverses analyses sont maintenant possibles sur cet ensemble de CV d'informaticiens. On peut par exemple comparer les niveaux de formation annoncés pour certains types de postes recherchés. Sur le diagramme ci-dessous, par exemple, on compare les trois postes de *chef de projet*, *administrateur système* et *responsable d'exploitation*.



De la même manière, on peut facilement regarder les corrélations entre certains postes et certaines compétences affichées. Sur le diagramme ci-dessous, on reprend les mêmes trois fonctions en montrant pour chacune les compétences les plus fréquemment citées, c'est à dire mises en valeur, dans les CV.



## Analyse des transitions

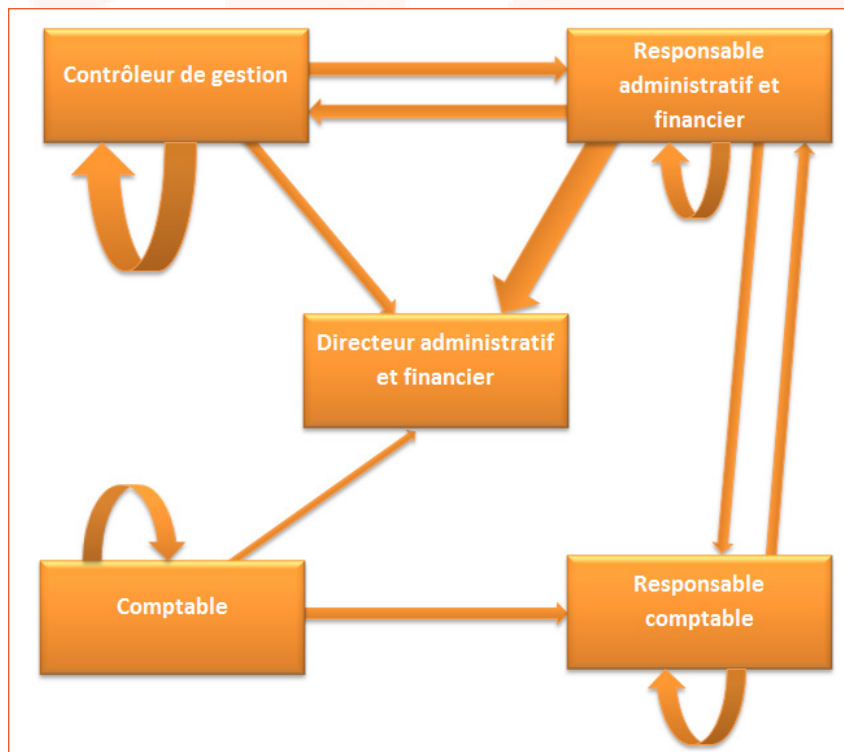
Dans tout CV, son auteur décrit les différents postes qu'il a eu au cours de sa carrière. L'analyseur est capable de reconnaître ces différents postes et de les replacer dans l'ordre chronologique. On a ainsi pour chaque CV, le poste actuel, et les deux postes précédents, quand ils existent.

A partir de ces informations, on peut construire des graphes de transition, qui montrent comment, en général, évoluent les carrières à travers une succession de postes.

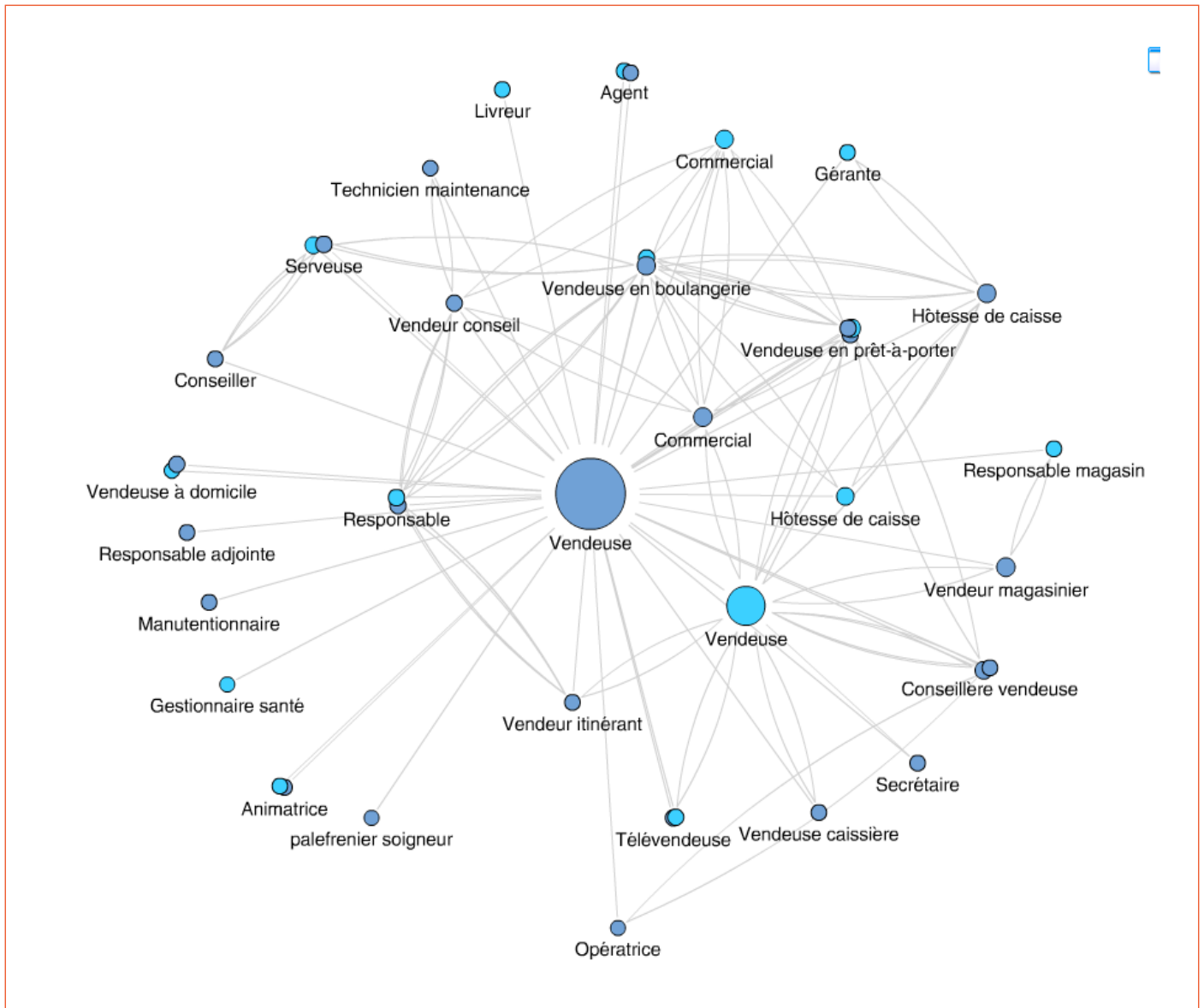
Dans l'exemple ci-dessous, on a analysé les transitions entre divers postes dans les fonctions financières et comptables. L'épaisseur des flèches est proportionnelle au nombre de transitions correspondantes.

On voit tout de suite qu'il y a beaucoup de stabilité d'un poste à l'autre : quand on change d'entreprise, on retrouve bien souvent le même intitulé de poste dans la nouvelle société, comme le montrent ci-dessous les flèches formant une boucle sur un même poste.

On voit ainsi que certaines relations sont très symétriques, montrant que l'on passe facilement d'une fonction à l'autre et réciproquement. C'est le cas par exemple entre « contrôleur de gestion » et « responsable administratif et financier ». Pour ceux qui sont familiers avec le monde de la gestion des entreprises, il est clair que ces deux fonctions sont proches, mais ne recouvrent pas exactement les mêmes types de missions. Il est amusant de voir que pour la fonction « administratif et financier » on passe facilement de « responsable » à « directeur », mais que l'inverse n'est pas vrai. Quand on a obtenu un titre de directeur, on le garde...



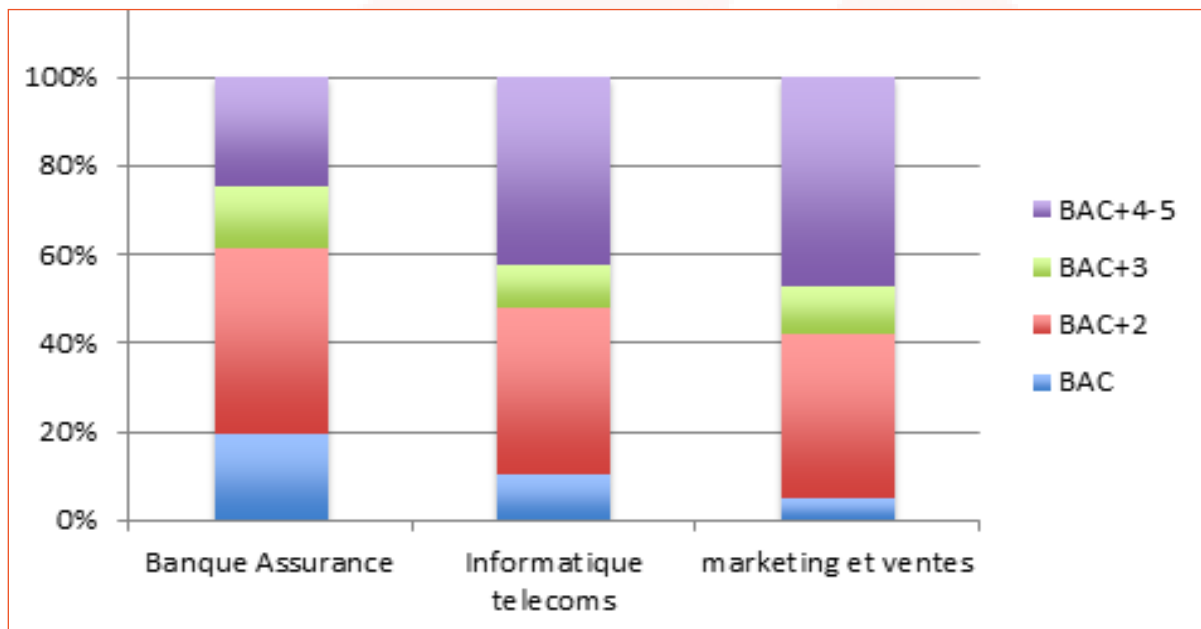
Un autre type de représentation permet une visualisation un peu différente. Sur le schéma ci-dessous, on représente en bleu foncé le poste actuel, et en bleu clair le poste précédent. Les surfaces sont proportionnelles au nombre de CV. Les arcs entre cercles représentent les transitions de postes.



## Analyse diplômes / branche professionnelle

L'analyse d'un grand nombre de CV permet également de rechercher des indications sur le niveau d'études par secteur d'activité, par type d'entreprise, voire par entreprise.

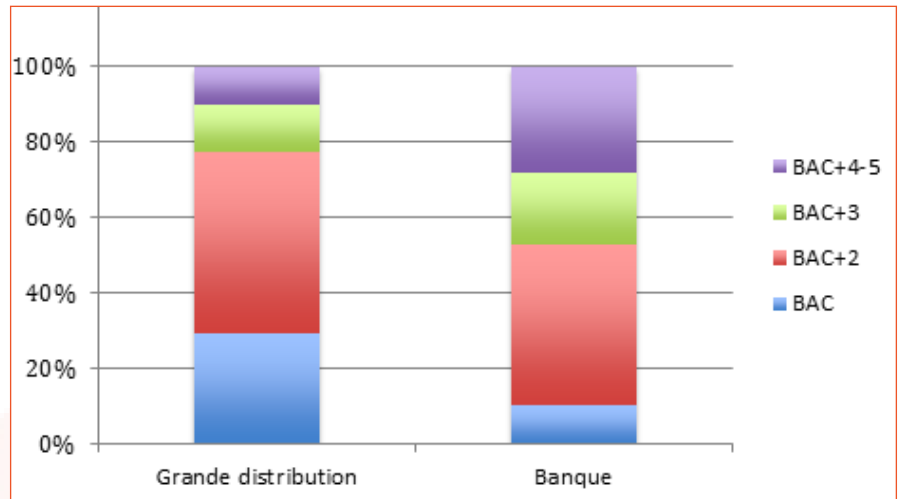
Sur le diagramme ci-dessous par exemple, on compare les niveaux de recrutements par secteur d'activité. Les données brutes ont été ramenées à des pourcentages pour les rendre comparables. On voit que le niveau BAC+3 est peu représenté, reflétant assez bien la structure de l'enseignement supérieur en France. On voit surtout, sur cet exemple qui compare trois secteurs d'activités différents, mais restant globalement dans le domaine des services, que les niveaux demandés sont très variables selon les secteurs. La Banque et Assurance recrute à tous les niveaux, alors que les métiers du marketing et de la vente sont les plus demandeurs de diplômes plus élevés.



La même analyse peut être faite au niveau beaucoup plus précis d'une entreprise, donnant ainsi un aperçu de ses besoins de recrutement.

Sur le diagramme ci-après, on compare deux grandes entreprises françaises, chacune majeure sur son marché. Leur identité n'est pas donnée ici pour des raisons de confidentialité des données utilisées, mais les volumes de CV citant ces deux entreprises sont similaires, environ 500 CV chacune.

On voit très clairement la différence entre les deux entreprises, celle du secteur de la grande distribution recrutant à un niveau moins avancé que la banque. On objectera que l'on montre là des évidences. Mais cela n'est pas si simple. Par exemple, on constate que la banque particulière analysée ci-contre a un profil de recrutement assez différent de l'ensemble du secteur Banque et Assurance comme vu précédemment, avec proportionnellement moins de recrutement au niveau Bac, au profit de formations plus avancées.



### Pourquoi ce type d'approche ?

Les résultats proposés ici ont surtout pour objet d'illustrer une méthode. L'échantillon de 40 000 CV environ utilisé n'est pas représentatif de l'ensemble du marché français, ni en termes de couverture géographique, ni en termes de types de métiers ou de fonctions.

Plusieurs types d'usage pourraient être imaginés : les job boards, qui disposent de grandes bases de CV sont les premiers à pouvoir les exploiter. Les grandes entreprises, ensuite, sont nombreuses à avoir leur propre CVthèque. Enfin, n'importe qui voulant faire ce type d'étude peut très facilement construire une base de CV par un simple crawl sur le web.

On peut envisager d'utiliser ces outils de text-mining dans différentes directions : pour définir et améliorer des programmes de formation par exemple, par l'analyse des transitions ou les corrélations entre diplômes et emplois occupés. Si l'on se rend compte par exemple qu'une proportion importante de CV montre qu'après un poste de secrétaire-comptable on peut devenir comptable dans une région donnée mais que ce n'est pas le cas dans une autre, peut-être faut-il chercher une explication dans un déficit de formation dans cette région.

Ce n'est là qu'un exemple fictif. Mais ce type d'analyse, popularisée avec la banalisation des outils de data-mining ou data-analytics trouve aujourd'hui un nouveau champ d'application avec l'essor du Big Data.