Estimation du nombre de citations de papillotes et de blagues Carambar

Philippe Gambette

$gambette@lirmm.fr$.
Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier.
C.N.R.S., Université Montpellier 2.
161 rue Ada, 34392 Montpellier Cedex 5 France.

8 juin 2009

RÉSUMÉ. *Les papillotes et les carambars sont deux gourmandises d'invention française dont l'intérêt principal est, pour de nombreux consommateurs peu gourmands, le papier qui les enrobe. Celui-ci contient une citation [1], un rébus, un dessin humoristique, une blague [2], ou plus récemment une création désopilante d'Élie Semoun* [1]*. Ces messages divers sont extraits d'un ensemble fini pour limiter les coûts de production. En supposant que la répartition des messages à l'intérieur d'un sachet de papillotes Révillon ou bonbons Carambar se fasse par tirage aléatoire (uniforme et indépendant), nous donnons une méthode pour estimer le nombre total de messages différents à partir d'un échantillon (par exemple, un sachet). Cette avancée fondée sur des calculs statistiques permet donc de résoudre un mystère essentiel sur la fabrication de ces gourmandises.*

MOTS-CLÉS : *Papillote, combinatoire, statistiques, chocolat, Carambar.*

1. Introduction

La *papillote* a été créée en 1790 à Lyon quand un apprenti du chocolatier Papillot a été surpris en train de dérober des chocolats qu'il envoyait entourés d'un billet doux à la demoiselle dont il était amoureux [1]. Son employeur l'a renvoyé en prenant soin de commercialiser son idée. Depuis, la papillote est devenue le chocolat traditionnel des fêtes de fin d'année [3], fabriquée notamment par l'entreprise Révillon Chocolatier, qui enrobe les papillotes de sa gamme "Festive" par des citations humoristiques ou philosophiques.

Chaque citation est présentée sur un petit papier qui en contient une entière, ainsi que des fractions d'une ou deux autres, à gauche et à droite. En observant la consécutivité des citations sur ces papiers, on se rend compte qu'elle est toujours maintenue, autrement dit que si une citation a apparaît immédiatement à droite d'une autre b sur un papier, elle apparaîtra également immédiatement à droite de b sur tout autre papier qui la contient, comme montré en Figure 1(a). On peut donc raisonnablement en déduire que ces papiers proviennent de l'impression puis du découpage d'un "dictionnaire" contenant un nombre fini de citations dans un ordre fixé.

Les mêmes observations peuvent être menées sur les blagues imprimées sur le papier enrobant les *Carambar*, illustrées en Figure 1(b). Ces friandises, actuellement commercialisées par Cadbury Schweppes, ont été créées en 1954 dans l'usine Delespaul-Havez. C'est en 1969 que les blagues ont fait leur apparition sur le papier d'emballage [4].

Nous nous intéressons donc à l'estimation de la taille de ces dictionnaires de citations ou de blagues à partir d'un échantillon (typiquement, un sachet). On estime, raisonnablement, que les citations présentes dans l'échantillon

1. `http://www.carambar.fr/html/elie_semoun.html`

(a) (b)

FIGURE 1. *Recollage de citations de papillotes Révillon (a) ou de blagues Carambar (b) chevauchantes.*

sont obtenues par un tirage aléatoire, uniforme (probabilités égales pour le tirage de chaque citation) et indépendant (la probabilité de choisir une citation et la probabilité de choisir la suivante sont indépendantes).

2. Estimation du maximum de vraisemblance

Nous choisissons de décrire un tirage de papillotes par le nombre d de citations différentes piochées, et d'estimer le nombre total de citations différentes par maximum de vraisemblance par rapport à la valeur observée de d.

Notons que cette formulation du problème passe par une discrétisation des données. En effet, les papiers contenant les citations ne sont pas découpés uniformément. Tous font apparaître une citation entière, ainsi qu'une portion ou la totalité de la citation qui la précède, et de celle qui la suit. Nous choisissons donc de représenter chaque papier d'emballage contenant une citation par le numéro d'identifiant de la citation qui est placée sur le point central du papier. Nous procédons de même pour les blagues Carambar qui présentent exactement le même problème.

On cherche donc à calculer la probabilité $P_{d,k}(n)$ de tirer d citations différentes parmi k piochées avec remise parmi un ensemble de papillotes où les n citations différentes sont également réparties.

On peut aisément définir $P_{d,k}(n)$ par récurrence :

$$\begin{aligned} P_{d,k}(n) &= P_{d-1,k-1}(n)\frac{n-d+1}{n} + P_{d,k-1}(n)\frac{d}{n} \text{ pour } 1 < d \leq n, k \in \mathbb{N}^*, \\ P_{1,k}(n) &= \frac{1}{n^{k-1}} \text{ pour } k, n \in \mathbb{N}^*, \\ P_{d,1}(n) &= 0 \text{ pour } 1 < d \leq n \in \mathbb{N}^*. \end{aligned}$$

Pour obtenir une formule plus directe facilitant les calculs, on peut remarquer que le problème est équivalent au dénombrement des mots de k lettres (choisies parmi un alphabet de n lettres) contenant exactement d lettres différentes. Appelons $a_{d,k}(n)$ ce nombre, on a donc :

$$P_{d,k}(n) = \frac{a_{d,k}(n)}{n^k}. \tag{1}$$

Remarquons à présent que pour calculer $a_{d,k}(n)$, il suffit de calculer le nombre $b_{d,k}$ mots de k lettres dont d différentes choisies parmi un alphabet de taille d, et multiplier par toutes les façons possibles de projeter ces d lettres à l'intérieur de l'alphabet de taille n. Ceci donne l'égalité :

$$a_{d,k}(n) = \binom{n}{d} b_{d,k}. \tag{2}$$

Comme $b_{d,k}$ ne dépend pas [1] de n, les égalités 1 et 2 suffisent pour calculer le maximum de $P_{d,k}(n)$ par rapport à n :

$$\max_n P_{d,k}(n) = \max_n \frac{\binom{n}{d}}{n^k}. \quad (3)$$

3. Résultats expérimentaux

3.1. *Estimation*

Nous avons effectué une dégustation de 52 papillotes Révillon de la gamme des papillotes “Festives”. Ce tirage a permis de déchiffrer (en utilisant le web pour compléter certaines citations tronquées) et affecter un identifiant à 65 citations. Le processus de discrétisation des données décrit en section 2 a conduit à trouver $d = 40$ citations différentes parmi les $k = 52$ tirées.

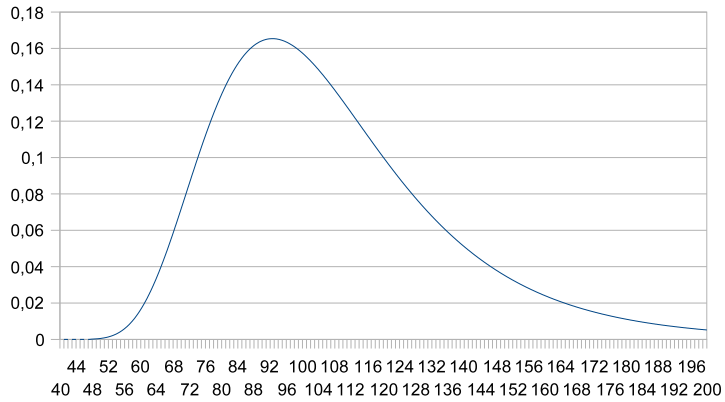

FIGURE 2. *Probabilité de tirer 40 citations différentes parmi 52, en fonction du nombre total de citations différentes.*

Nous présentons en Figure 2 la courbe de probabilité du nombre de citations différentes de ce tirage en fonction du nombre total de citations différentes. Cette courbe atteint son maximum sur $\mathbb{N}$ pour $n = 93$, avec une probabilité de 16.5%.

3.2. *Précision*

Pour évaluer la précision de ce résultat, nous créons plusieurs jeux de données artificiels par un rééchantillonage de type Jack-knife, c’est à dire un tirage aléatoire de 45 citations parmi les 52 réellement tirées, et nous effectuons les mêmes calculs, dont les résultats sont présentés dans la Table 1. Ceux-ci permettent de fournir une estimation moyenne de $n = 83$ et un intervalle de confiance de [74,108]. On peut donc s’attendre à une erreur de 30%.

Tirage	1	2	3	4	5	6	7	8	9	10
$d =$	35	35	35	34	35	37	35	34	34	35
$\arg\max_n P_{d,k}(n) =$	84	84	84	74	84	108	84	74	74	84

TABLE 1. *Résultats de l’estimation du nombre de citations sur 10 tirages aléatoires de 45 citations parmi 52.*

1. Le calcul de $b_{d,k}$ est détaillé sur `http://www.physicsforums.com/showthread.php?t=301013`.

En fait, répéter une expérience similaire avec un tirage aléatoire de 25 citations (voir Table 2) permet de montrer les limites de la méthode. En effet, les valeurs possibles de n trouvées par maximum de vraisemblance arrivent dans l'intervalle [34,92], et conduisent à une estimation moyenne de 56 citations, alors que l'on sait qu'il y a au moins 65 citations différentes.

Tirage	1	2	3	4	5	6	7	8	9	10
$d =$	22	22	19	19	18	19	19	22	20	18
$\arg\max_n P_{d,k}(n) =$	92	92	41	41	34	41	41	92	52	34

TABLE 2. *Résultats de l'estimation du nombre de citations sur 10 tirages aléatoires de 25 citations parmi 52.*

L'application de ces petits tests pour donner une idée de la précision des données nous semble donc indispensable afin d'éviter de fournir des estimations trop éloignées de la réalité.

3.3. *Application aux données Carambar*

Cette méthode d'estimation a été utilisée sur les résultats obtenus progressivement à partir de tirages successifs, pour fournir les résultats présentés en Figure 3. Toutefois, ces résultats sont sous-estimés. En effet, la blague du 56ème Carambar dégusté a permis de constater que contrairement à notre hypothèse initiale, une blague n'est pas toujours précédée et suivie des mêmes. Ainsi, il est possible que certaines blagues sur-représentées conduisent à des erreurs d'estimation. En outre, on peut remarquer que contrairement aux papillotes de largeur constante, les blagues Carambar, plus ou moins élaborées, n'ont pas une hauteur constante, ce qui conduit à un tirage favorisé des blagues plus longues. Finalement, ces problèmes conduisent probablement à une sous-estimation du nombre total de blagues.

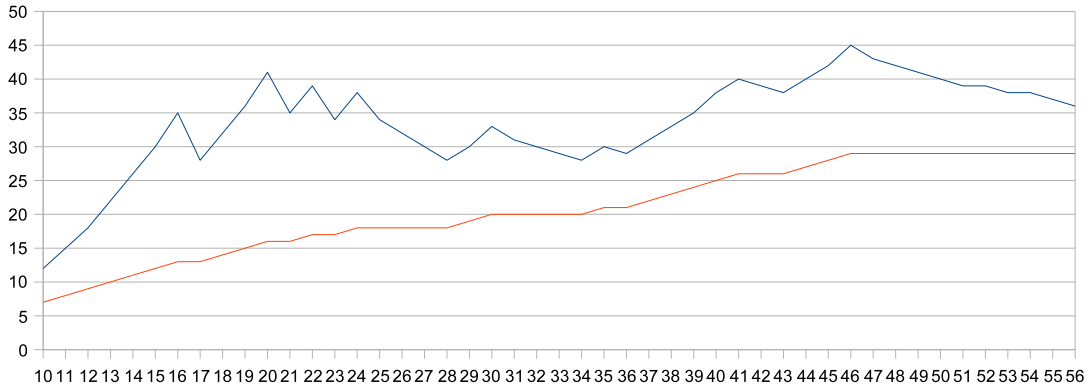

FIGURE 3. *Evolution de l'estimation du nombre total de blagues Carambar (courbe bleue) en fonction de la taille du tirage, et du nombre de blagues différentes tirées (courbe rouge).*

4. Conclusion

Nous avons contacté l'entreprise Révillon Chocolatier [2] qui nous a aimablement dévoilé que le nombre total de citations pour la gamme des papillotes "Festives" était 108. Notre estimation directe de 93 citations correspond à une erreur de 13,9%, soit un ordre de grandeur tout à fait satisfaisant.

2. `http://www.papillotesrevillon.fr/`

La précision atteinte par la méthode présentée ici semble améliorable, vraisemblablement en choisissant un autre paramètre caractéristique du tirage pour l'étude du maximum de vraisemblance : la taille de la plus longue séquence de citations consécutives, le nombre de citations présentes 2 fois, la distribution des nombres d'apparitions de citations...

Nous cherchons aussi à appliquer cette méthode sur d'autres données, comme sur celles de suivi des billets en euros du site EuroBillTracker [3], afin de vérifier si l'on obtient une bonne estimation du nombre total de billets en euros en circulation (11,8 milliards fin 2008 d'après la Banque Centrale Européenne). Ceci conduit à un problème de calcul efficace avec de grands entiers, et incite à trouver une formule directe d'estimation du maximum de vraisemblance.

5. Bibliographie

[1] Collectif L'inventaire Du Patrimoine Culinaire De La France. *Rhône-Alpes - Produits Du Terroir Et Recettes Traditionnelles*. Albin Michel / CNAC - région Rhône-Alpes (1995).

[2] Collectif. *Les Blagues Carambar*. Éditions Michel Lafon (2004).

[3] Brigitte Brégeon-Poli. "Va pour treize !" La "tradition" des desserts de Noël en Provence. *Terrain* 24 (1995), pp. 145-152.

[4] Cadbury Schweppes. *Le dossier de marque Carambar*. `http://www.carambar.fr/download/dossiers/Dossier_de_marque_carambar.pdf` (2008).

3. `http://www.eurobilltracker.com`