رتبه بندی لغات و آنتروپی کسری

مهری‌دهنوی، حسین[۱]؛ آگاهی، حمزه[۲]؛ مهری، علی[۱]؛ تابش، مهسا[۱]

[۱] گروه فیزیک، دانشگاه صنعتی نوشیروانی، خیابان شریعتی، ۷۱۱۶۷-۴۷۱۴۸، بابل

[۲] گروه ریاضی، دانشگاه صنعتی نوشیروانی، خیابان شریعتی، ۷۱۱۶۷-۴۷۱۴۸، بابل

چکیده

در این مقاله، روشی جدید در متن‌کاوی را با استفاده از آنتروپی کسری ارائه می‌دهیم. نتایج این روش برای استخراج نمایه‌ی کتاب آماری کسلا و برگر (۱۹۹۰) نشان می‌دهد که آنتروپی کسری، نسبت به آنتروپی متداول شنون، ابزار موفق‌تری در متن‌کاوی است.

Word Ranking and fractional entropy

Mehri-Dehnavi, Hossein[1]; Agahi, hamzeh[2]; Mehri, Ali[1]; Tabesh, Mahsa[1]

[1] Department of Physics, Babol Noshirvani University of Technology, Shariati Ave., Babol, 47148-71167, Iran,

[2]Department of Mathematics, Babol Noshirvani University of Technology, Shariati Ave., Babol, 47148-71167, Iran,

In this paper, we propose a new method in text mining using the fractional entropy. Using this method for keyword extraction of the Statistical Inference book by Casella and Berger (1990) indicates that fractional entropy is more efficient than the traditional Shannon entropy in text mining.

PACS No. 05

مقدمه

یک مفهوم حائز اهمیت در آمار و احتمالات و فیزیک آماری آنتروپی شنون نام دارد. همان‌طور که از نام این مفهوم برداشت می‌شود، این مفهوم نشان‌دهنده‌ی میزان بی‌نظمی سیستم مورد مطالعه است. آنتروپی شنون به صورت زیر تعریف می‌شود

$$S=\sum_i p_i\, I(p_i)=-\sum_i p_i \ln p_i\ , \qquad (۱)$$

که $I(p_i)=-\ln p_i$ اطلاعات شنون و p_i احتمال حضور سیستم مورد مطالعه در حالت iام می‌باشد، و در شرط $\sum_i p_i=1$ صدق می‌کند.

یک شاخه‌ی مهم ازریاضیات که اخیرا مورد توجه محققین قرار گرفته حسابان با مشتقات و انتگرالهای کسری است [۲]. با استفاده از مفاهیم مشتق و انتگرال کسری، تعمیمی از آنتروپی شنون تحت عنوان آنتروپی کسری ارائه شد [٤]. آنتروپی کسری از مرتبه‌ی $\alpha\in R$ به صورت زیر تعریف می‌شود

$$S_\alpha=\sum_i\{-\frac{p_i^{-\alpha}}{\Gamma(1-\alpha)}[\ln p_i+\psi(1)-\psi(1-\alpha)]\}p_i. \qquad (۲)$$

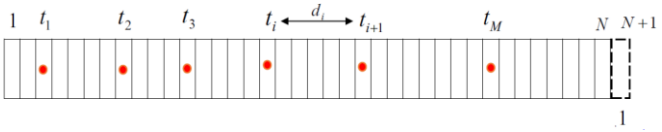
$\Gamma(x)$ و $\psi(x)$ به ترتیب نشان‌دهنده‌ی توابع گاما و دی‌گاما می‌باشند. آنتروپی کسری تعمیمی از آنتروپی شنون می‌باشد که در حالت خاص $\alpha=0$ بر آنتروپی شنون منطبق می‌شود. کاربردهایی از آنتروپی کسری در منابع [۱و۷] ارائه شده است.

در بخش بعدی، پس از مروری بر روش متداول متن‌کاوی بر اساس آنتروپی شنون، به ارائه مفهوم اهمیت تعمیم‌یافته کسری لغات، WI_α، با استفاده از آنتروپی کسری، پرداخته و بر اساس آن به ارائه روش جدیدی برای متن‌کاوی بر اساس آنتروپی کسری خواهیم پرداخت. در بخش سوم روش ارائه شد خود برای استخراج لغات مهم کتاب آمار استنباطی استفاده می‌نماییم [۳].

۲ متن کاوی

احتمال در متن کاوی

مطابق با شکل (۱) متنی به طول N را در نظر بگیرید. با این توصیف متن مورد بررسی ما می‌تواند شامل N لغت باشد. فرض بر این است که لغات متن مطابق شکل از ابتدا تا به انتها رتبه‌بندی شده‌اند. همچنین طبق معمول فرض دوره‌ای بودن متن را نیز به کار می‌بریم. کلمه‌ی (مورد نظر) W_1 را در نظر بگیرید که به تعداد (فرکانس) M بار در متن مورد مطالعه تکرار شده باشد. شماره و جایگاه‌های مختلف این کلمه در متن را مطابق با شکل به ترتیب با $t_1, t_2, ..., t_M$ نشان می‌دهیم. با توجه به توضیحات فوق فاصله‌ی کلمه (مورد نظر) W_1 شماره i با کلمه مشابه بعدی خود در متن برابر با $d_i = t_{i+1} - t_i$, خواهد بود.



شکل۱: تصویر نمادین از یک کتاب با طول N کلمه و لغت مشخص W_1 با تعداد تکرار M مرتبه که در جایگاههای t_i قرار گرفته است. متن به طور دورهای فرض شده است (یعنی جایگاه کلمه‌ی $N+1$ بر جایگاه کلمه‌ی اول منطبق است).

برای محاسبه‌ی فاصله‌ی آخرین کلمه‌ی متن با کلمه‌ی بعدی خود باید فرض چرخه‌ای را در نظر بگیریم. به عبارتی با فرض $t_{M+1} \to t_1 + N$ خواهیم داشت:

$$d_M = t_{M+1} - t_M = t_1 - t_M + N. \qquad (۳)$$

واضح است که جمع فاصله‌های تمامی کلمات برابر با طول کل متن می‌باشد، به عبارتی $\sum_i^M t_i = N$. حال می‌توانیم برای کلمه مورد نظر W_1 قرار گرفته در جایگاه t_i، یک احتمال به صورت $p_i = \frac{d_i}{N}$، تعریف نمود. واضح است که جمع احتمال‌های مربوط به یک لغت مورد نظر برابر با یک می‌باشد. روش‌های دیگری هم برای تعریف احتمال مربوط به یک کلمه در متن وجود دارد که تعدادی از آن‌ها در[۵و۶] مرور شده‌اند.

روش متداول رتبه‌بندی لغات با آنتروپی شنون

در روش متداول رتبه‌بندی لغات یک متن مطابق فرایند فوق برای هر کلمه‌ای یک مجموعه احتمالات $\{p_1, p_2, ... p_M\}$ را برای کلمه‌ی دلخواه W_j محاسبه می‌کنند، سپس با استفاده از رابطه‌ی (۱) آنتروپی شنون مربوط به آن کلمه را به دست می‌آورند. حال اگر آن کلمه موردنظر یک لغت بی‌اهمیت مانند "است"، "در" و ... باشد، این کلمه تقریباً در تمامی متن به طور یکنواخت توزیع شده است. ولی اگر کلمه‌ی مربوطه کلمه‌ای مانند "احتمال"، "توزیع"، "همبستگی" و ... باشد، واضح است که این کلمه در بخش‌های خاصی از متن از قبیل بحث روی نتایج، و نمودارها و ... در مقایسه با قسمت مقدمات متن بیشتر تکرار می‌شود.

اگر در حالت ایده‌آل یک لغت بی‌اهمیت W_1 با تکرار M بار در متن به طور کاملاً یکنواخت توزیع شده باشد. واضح است که $p_i = d_i/N = \frac{1}{M}, d_i = N/M$، و مقدار آنتروپی شنون مربوطه برابر با $S_{uniform} = \sum_{i=1}^{M} \frac{1}{M} \ln M = \ln M$، خواهد بود. که بیشترین مقدار برای یک کلمه با تعداد تکرار M عدد در متن است. ولی اگر کلمه‌ی موردنظر کلمه‌ای مهم از تعداد تکرار M مرتبه در متن باشد، واضح است که آنتروپی مربوطه از مقدار $\ln M$ کمتر خواهد شد. به همین دلیل هر چقدر مقدار آنتروپی یک لغت با مرتبه تکرار M از مقدار آنتروپی لغتی با همان فرکانس و توزیع یکنواخت $(S^M_{uniform} = \ln M)$ متفاوت‌تر باشد این کلمه، کلمه‌ی مهم‌تری است. این روش، روش متداول رتبه‌بندی لغات با آنتروپی شنون است. پس به نظر می‌آید که اختلاف آنتروپی یک کلمه با فرکانس مشخص M و آنتروپی کلمه‌ای با توزیع یکنواخت و همان فرکانس معیار خوبی برای تشخیص اهمیت کلمه‌ی مربوطه باشد. این امر تا حدی درست می‌باشد! دلیل این امر نرمالیزه نبودن معیار اشاره شده است. برای نرمالیزه نمودن این معیار بهتر است اختلاف به دست آمده را بر مقدار آنتروپی لغت با توزیع یکنواخت از همان فرکانس، $S^M_{uniform} = \ln M$، تقسیم نماییم. این معیار را اهمیت لغت می‌نامیم.

تعریف ۱. اهمیت یک لغت با مرتبه تکرار M در یک متن به صورت زیر تعریف می‌شود

$$WI(W_1) = \frac{\left| S^M_{uniform} - S(W_1) \right|}{S^M_{uniform}}, \qquad (۴)$$

که در آن $S(W_1)$، و $S^M_{uniform} = \ln M$ به ترتیب آنتروپی لغت W_1 با تکرار M، مرتبه و آنتروپی یک لغت با توزیع یکنواخت همان مرتبه تکرار می‌باشند.

معیارهای ارزیابی روش‌ها در متن‌کاوی

در این بخش به مرور معیارهای متداول برای متن‌کاوی می‌پردازیم. متنی را در نظر بگیرید که لغات مهم آن از قبل نمایه شده باشند. تعداد لغات حاضر در نمایه این کتاب را N_{rel} فرض می‌کنیم. پس اجرای برنامه و امتیازبندی لغات طبق رابطه‌ی (٤)، به لیستی از لغات که بر حسب معیار اهمیت لغت مرتب شده‌اند خواهیم رسید. تعداد N_{ret} لغت اول به دست آمده در این لیست را در نظر بگیرید. از این تعداد، تعداد $N_{rel\cap ret}$ آن لغاتی هستند که نمایه متن موردنظر نیز وجود دارند. نسبت این دو عبارت را معیار "صحت"[۱] الگوریتم می‌نامیم:

$$P=\frac{N_{rel\cap ret}}{N_{ret}} \qquad (۵)$$

همچنین نسبت تعداد لغت مشترک به دست آمده در هر مرحله با نمایه، $N_{rel\cap ret}$ به تعداد کل لغات لیست نمایه را هم به عنوان معیار " بازیابی"[۲] تعریف می‌کنند:

$$R=\frac{N_{rel\cap ret}}{N_{rel}} \qquad (۶)$$

اگر N_{ret} به عنوان متغیر فرض کنیم و معیارهای صحت و بازیابی را بر حسب این متغیر رسم کنیم، می‌توانیم رفتار این توابع را برحسب متغیر ذکر شده بررسی کنیم. با توجه به تعریف ارائه شده برای این معیارها، رفتار هر دو آن‌ها برای مقادیر کوچک متغیر N_{ret} رفتار نوسانی دارند. همچنین برای مقادیر بزرگ N_{ret} رفتار معیارهای بازیابی و صحت به ترتیب صعودی و نزولی هستند.

معیار دیگری که میانگین هارمونیک دو معیار صحت و بازیابی می‌باشد را "معیار- F"[۳] می‌نامند:

$$F=\frac{2RP}{R+P} \qquad (۷)$$

رفتار کلی (به جز نقاط اولیه نوسانی) این معیار بر حسب متغیر N_{ret}، با دو معیار بالا متفاوت است. این معیار ابتدا رفتار صعودی داشته و پس از نقطه‌ی بیشینه‌ی خود رفتار نزولی خواهد داشت.

---

[۱] precision

[۲] Recall

[۳] F-masure

برای هر الگوریتم داده‌کاوی و یا متن‌کاوی، مقدار بیشینه‌ی معیار- F عددی بخصوص برای متن مورد کاوش قرار گرفته است. الگوریتمی که مقدار بیشینه‌ی معیار- F آن، عدد بزرگ‌تری را برای متن خاصی به خود بگیرد، آن الگوریتم موفق‌تری در بررسی آن متن می‌باشد. اخیراً کاربردهای وسیعی از این آنتروپی در رتبه‌بندی کلمات و متن‌کاوی ارائه شده است [۵].

روش پیشنهادی برای رتبه‌بندی لغات با استفاده از آنتروپی کسری

در زیر بخش قبلی روشی را بیان نمودیم که برای هر لغت موردنظر W_1 می‌توان مجموعه احتمال تعریف نمود، و آنگاه با مجموعه احتمال به دست آمده و با استفاده از رابطه‌ی (۱) آنتروپی شنون مربوط به آن لغت را محاسبه نمود. همچنین نشان دادیم که لغات کم اهمیت آنتروپی بیشتری خواهند داشت. همانند روش اشاره شده در زیر بخش قبلی می‌توان احتمال‌های مربوط به یک لغت در یک متن را به دست آورد. پیشنهاد ما این است که برای به دست آوردن لغات مهم یک متن به جای آنتروپی شنون (۱) از آنتروپی کسری S^{α} ارائه شده در رابطه‌ی (۲) استفاده کنیم. نتایج حاصل از روش متن‌کاوی با آنتروپی کسری، برای حالت خاص $\alpha=0$، برابر با نتایج به دست آمده برای متن‌کاوی با استفاده از آنتروپی شنون خواهد بود.

تعریف ۲ . اهمیت تعمیم‌یافته کسری یک لغت با مرتبه تکرار M، در یک متن به صورت زیر تعریف می‌شود

$$WI_{\alpha}(W_1)=\frac{\left|S^{M}_{\alpha,uniform}-S_{\alpha}(W_1)\right|}{S^{M}_{\alpha,uniform}}, \qquad (۸)$$

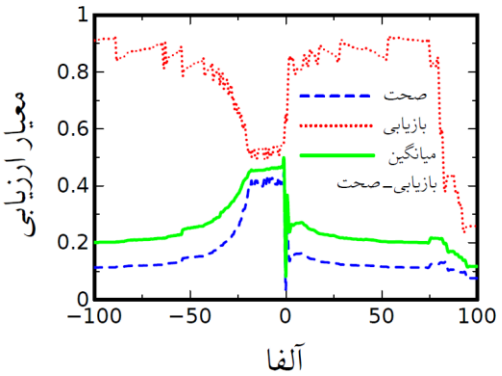
که در آن $S_{\alpha}(W_1)$، و $S^{M}_{\alpha,uniform}$، به ترتیب آنتروپی کسری مرتبه‌ی α لغت W_1 با تکرار M مرتبه، و آنتروپی کسری با مرتبه‌ی α یک لغت با همان مرتبه تکرار و توزیع یکنواخت، می‌باشند.

۳ نتایج مدل ارائه شده برای کتاب کسلا و برگر

در این بخش به بررسی روش پیشنهادی برای رتبه‌بندی کلمات در کتاب آماری کسلا و برگر خواهیم پرداخت. این مرجع شامل

$N = ۲۱۸۱۳۳$ کلمه است. از این تعداد کلمه، تعداد کل $N_v = ۱۱۶۶۵$ کلمه متفاوت وجود دارد.

در این مقاله ابتدا طبق تعریف ۲ لغات کتاب را بر حسب اهمیت تعمیم‌یافته کسری لغات برای یک α مشخص رتبه‌بندی کردیم. برای یک مقدار مشخص آلفا، با استفاده از نمایه‌ی کتاب و روش پیشنهادی در بخش قبلی می‌توانیم نمودار صحت، بازیابی، و همچنین هارمونیک این دو معیار را بر حسب تعداد لغت ظاهر شده در لیست لغات مرتب شده بر اساس اهمیت لغات، به دست آوریم. سپس بیشینه مقدار معیار- F را برای پارامتر کسری α موردنظر به دست می‌آوریم. مقادیر صحت و بازیابی برای مربوط به نقطه‌ی بهینه‌ی معیار- F را نیز می‌توانیم محاسبه کنیم. شکل ۲ نتایج به دست آمده، را برحسب مقادیر مختلف پارامتر کسری را رسم نموده است. همان‌طور که در شکل مشاهده می‌نماییم، بیشترین مقدار میانگین هارمونیک بازیابی و صحت در $\alpha = -۰/۸$ رخ می‌دهد.



شکل ۲ : بازی‌ابی (خط‌چین آبی)، صحت (نقطه‌چین قرمز) و می‌انگین آن‌ها (خط سبز) برای نمای‌ه‌های استخراج شده از کتاب آماری کسلا و برگر(۱۹۹۰) به کمک آنتروپی کسری با آلفاهای گوناگون. بیشترین مقدار می‌انگین هارمونی‌ک بازی‌ابی و صحت در WI ـ۰/۸ رخ می‌دهد.

بحث و نتیجه‌گیری

با استفاده از آنتروپی کسری به معرفی اهمیت تعمیم‌یافته کسری لغات از مرتبه‌ی α، WI_α، پرداختیم، که در حالت خاص $\alpha = ۰$، بااهمیت لغات به دست آمده از آنتروپی متداول شنون، WI برابر است. سپس با استفاده از این معیار به استخراج لغات بااهمیت مرجع [۳]، پرداختیم. نتایج حاصل نشان دادند که مقدار بهینه پارامتر برای آنتروپی کسری برای متن‌کاوی کتاب مورد مطالعه برابر با $\alpha = -۰/۸$، می‌باشد.

جدول ۱ : مقای‌سه کمی نتای‌ج حاصل از آنتروپی کسری و آنتروپی شنون در استخراج نمای‌ه برای کتاب آمار توصی‌فی کسلا-برگر. ردی‌ف اول نشان‌دهنده‌ی نتای‌ج به دست آمده با استفاده از اهمی‌ت لغات تعمی‌م‌ی‌افته WI ـ۰/۸، ردی‌ف دوم نتای‌ج به دست آمده با استفاده از اهمی‌ت لغات WI به دست آمدند.

	جمعی‌ت نسبی نمای‌ه	بازی‌ابی	صحت	می‌انگی‌ن هارمونی‌ک بازی‌ابی و صحت
آنتروپی کسری	۰/۰۵	۰/۴۵	۰/۵۴	۰/۴۹
آنروپی شنون	۰/۵۶	۰/۰۵	۰/۶۴	۰/۰۹

با توجه به بررسی‌های انجام شده در این پژوهش روش رتبه‌بندی کلمات پیشنهادی از روش رتبه‌بندی با استفاده از آنتروپی شنون، در جداسازی لغات مهم یک متن بهتر عمل می‌کند. همچنین جدول ۱ به صورت کمی نیز نشان که آنتروپی کسری، نسبت به آنتروپی متداول شنون، ابزار موفق‌تر و سریع‌تری در متن‌کاوی است.

مراجع

[۱] Bagci, G. B. (2016). The third law of thermodynamics and the fractional entropies, *Physics Letters A*, 380(34), 2615-2618.

[۲] Baeanu, D., Diethelm, K.., Scalas, E., Trujillo, J.J. (2012). *Fractional Calculus*, world Scientific, Singapore.

[۳] Casella, G. Berger, R. L. (1990). *Statistical Inference*, Wadsworth, California.

[٤] Machado, J.T. (2014). Fractional order generalized information, *Entropy*, 16(4), 2350-2361.

[۵] Mehri, A. Darooneh, A.H. (2011-a). Keyword extraction by nonextensivity measure, *Physical Review* E 83, 056106.

[۶] Mehri, A. Darooneh, A.H. (2011-b). The role of entropy in word ranking, *Physica A*, 390, 3157-3163.

[۷] Lopes, A. M., Machado, J.A.T. (2016). Integer and fractional-order entropy analysis of earthquake data series, *Nonlinear Dynamics*, 84 (1), 79-90.