

مه داده

کریس باد

برگردان: علی محمدیان مصمم*، ویراستار: حسن حقیقی**

منبع اصلی این اطلاعات، محتویات در حال رشد وب سایت‌های شبکه‌های اجتماعی است. به عنوان مثال، فیس‌بوک^۲ در سال ۲۰۰۴ راه اندازی شد و در حال حاضر ۲ میلیارد کاربر ثبت شده دارد (حدود یک چهارم جمعیت کل جهان) که ۱/۵ میلیارد نفر از آن‌ها فعال هستند. هر روز حدود ۲/۵ میلیارد قطعه محتوی (حدود ۵۰۰ ترابایت اطلاعات) به فیس‌بوک اضافه می‌شود، که قسمت اعظم این داده‌های ذخیره شده به فرمت تصویری هستند. برآورد می‌شود که برنامه جستجوگر گوگل در میان حدود ۱۰^{۱۵} بایت داده (که این جستجو را با استفاده از الگوریتم‌های هوشمند ریاضی انجام می‌دهد) به جستجوی اطلاعات می‌پردازد.

منبع دیگر تولید مه داده‌ها، گوشی‌های همراه و تلفن‌های هوشمند^۳ هستند. در حال حاضر تعداد گوشی‌های موبایل از تعداد جمعیت جهان بیشتر است و به‌طور بالقوه امکان ۲۵,۰۰۰,۰۰۰,۰۰۰,۰۰۰,۰۰۰ ارتباط همزمان را دارند. برنامه‌های آینده برای یک شبکه ۵G در نظر دارد داده‌ها را به‌طور همزمان با نرخ برابر یک گیگا بایت در ثانیه در اختیار ده‌ها کاربر واقع در یک طبقه از یک سازمان قرار دهد.

از دیگر فن‌آوری در حال توسعه سریع، حسگرها^۴ هستند که می‌توانند مثلاً بطور مداوم وضعیت سلامتی ما را (و با رعایت الزامات اخلاقی مترتب بر آن) مانیتور کنند. شبکه ۵G قادر است چند صد هزار ارتباط همزمان شبکه‌های عظیمی از حسگرها را پشتیبانی نماید. در واقع آینده به سرعت در حال نزدیک شدن است: به زودی وسایل مورد استفاده ما به سادگی با کمترین دخالتی یا حتی بدون دخالت انسان با یکدیگر به مبادله اطلاعات خواهند پرداخت (به عنوان مثال در هر بار آماده سازی غذا، اجاق آشپزی با ماشین ظرفشویی و همچنین سوپر مارکت به مبادله اطلاعات خواهند پرداخت) - این وضعیت، اینترنت اشیاء^۵ نامیده می‌شود.

مقدار معتناهی از داده‌ها، قابل توجه از نظر علوم اجتماعی، نیز از روشی که ما از دستگاه‌ها استفاده می‌کنیم حاصل می‌شوند و اطلاعاتی درباره شیوه زندگی ما به دست می‌دهند. دوباره مباحث اخلاقی مهمی در اینجا وجود دارد. هر بار که ما از سایت آمازون خرید می‌کنیم،

این متن ترجمه‌ای از مقاله زیر است:

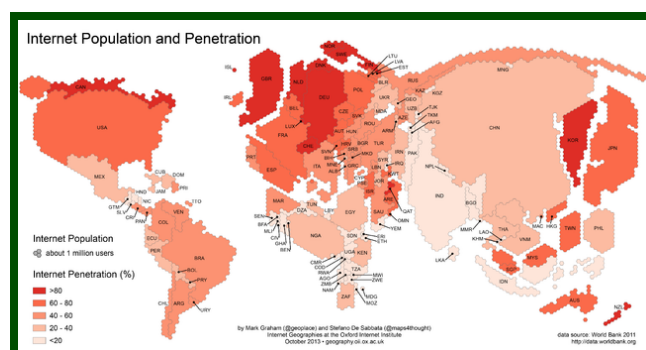
Chris Budd, Big Data, Plus Magazine, 2016.

<https://plus.maths.org/content/big-data>

چکیده: ما در عصر اطلاعات زندگی می‌کنیم. بیشتر کارهایی که انجام می‌دهیم، شدیداً متأثر از دسترسی ما به حجم انبوهی از داده‌ها است، خواه این کار از طریق اینترنت، بروی کامپیوترهای ما صورت بگیرد، یا بروی گوشی‌های تلفن همراه. کلمه نادقیق برای توصیف این سیل از اطلاعات، مه داده^۱ است. در سال ۲۰۱۲ دولت انگلستان مه داده را به عنوان یکی از هشت تکنولوژی برتر آینده معرفی کرد. حال چالش‌های مه داده چیستند و ما چگونه با آن مواجه می‌شویم؟

منبع تولید مه داده‌ها کجاست؟

شاید عمده‌ترین منبع فعلی مه داده اینترنت است. با توجه به برآوردهای اخیر هر ساله حدود ۱۰^{۲۱} بایت (یک زتا بایت) اطلاعات، که بسیاری از آن به فرمت گرافیکی است، به اینترنت اضافه می‌شود. نفوذ اینترنت در بریتانیا بیش از ۸۰ درصد و در اغلب کشورها به استثنای تعداد محدودی، بالای ۲۰ درصد است.



شکل ۱: این نقشه تعداد کاربران اینترنت در هر کشور و همچنین درصد جمعیتی که در سال ۲۰۱۱ به اینترنت دسترسی داشتند را نشان می‌دهد. تصویر از استفانو دو ساباتا و مارک گراهام

^۱Big Data ^۲Facebook ^۳smart phones ^۴sensors ^۵Internet of things

چه سؤالاتی باید درباره‌ی مه‌داده پرسیم؟

چگونه تصویری از مه‌داده بسازیم، درمورد آنها گمانه زنی کنیم، آنها را مدل سازی کنیم، و آنها را بفهمیم؟ ما چگونه در سیستم‌هایی که چنین داده‌هایی را تولید می‌کنند تجربه و آزمایش انجام دهیم و در نهایت چگونه آن سیستم‌ها را تحت کنترل قرار دهیم؟ چالش‌های ریاضی و علمی که در پس این سؤالات هستند هم متنوع و هم مهم و در مقیاس مه‌داده، اتوماسیون را الزامی می‌کند. این اتوماسیون به نوبه خود متکی بر الگوریتم‌های ریاضی است.

سؤالاتی که می‌توانیم در مورد مه‌داده پرسیم عبارتند از:

- چگونه اطلاعات حاصل از شبکه‌های گسترده در مرورگرهای وب مانند Google را رتبه‌بندی کنیم؟
- چگونه اولویت‌ها، وابستگی‌ها و یا حتی تمایلات مصرف‌کنندگان این داده‌ها را تشخیص دهیم و چگونه به آنها توصیه‌های ویژه آنها ارائه دهیم؟
- چگونه عدم اطمینان^۷ در روند درمانی هر بیمار خاص را مدل سازی کنیم؟
- چگونه می‌توانیم به مانیتورینگ برخط سلامتی فرد (به خصوص در محیطی که شبکه ۵G به آن هدایت خواهد شد) دست یابیم؟

- چگونه از داده‌های هوشمند برای عرضه انرژی استفاده کنیم؟
- من فکر می‌کنم نسبتاً درست خواهد بود که بگوییم بسیاری از پیشرفت‌های آتی در ریاضیات مدرن (به‌همراه علوم نظری رایانه) توسط کاربردهای مه‌داده برانگیخته خواهد شد یا توسط نیاز به فهم مه‌داده راهبری خواهد شد. بسیاری از تکنیک‌های ریاضی موجود (برخی از آنها که تا همین اواخر به عنوان ریاضیات محض در نظر گرفته می‌شدند) کاربردهای قابل توجهی در فهم ما از مه‌داده پیدا می‌کنند. یک نمونه کلیدی از این، ریاضیات نظریه شبکه^۸ است.

همه جا شبکه

نظریه شبکه، همانطور که از نام آن پیداست، اشیایی موسوم به گره‌ها^۹، که توسط یال‌ها^{۱۰} به یکدیگر پیوند داده شده‌اند را توصیف می‌کنند. گره‌ها می‌توانند رایانه‌ها یا وب سایت‌ها باشند و یال‌ها می‌توانند اتصال بین رایانه‌ها یا پیوند بین وب سایت‌ها باشند. همچنین گره‌ها می‌توانند مردم و پیوندها می‌توانند دوستان آنها در فیس‌بوک یا توییتر^{۱۱} باشند، و یا آنها می‌توانند تلفن همراه و مکالمات و هرگونه

از بانک اینترنتی استفاده می‌کنیم، دستگاه‌های الکتریکی را روشن می‌کنیم، یا به سادگی از تلفن همراه استفاده می‌کنیم یا یک ایمیل می‌فرستیم، در واقع درحال تولید داده هستیم که حاوی اطلاعاتی است که در اصل می‌توانند تجزیه و تحلیل شوند. به عنوان مثال، عادات خریدمان می‌تواند مشخص شود، یا محل‌های رفت و آمد ما می‌تواند ردیابی و ثبت شود. ریاضیات می‌تواند در تمام این مراحل مورد استفاده قرار گیرد اما ما هیچوقت نباید بعد اخلاقی این کار را از یاد ببریم.



شکل ۲: ساعت هوشمند^۶ می‌تواند اطلاعات به روز سلامتی شما را در هر جا که هستید به شما بدهد

ماهیت مه‌داده

به یک معنا، مه‌داده، موضوع تحقیقات ریاضی در طی حداقل ۱۰۰ سال گذشته بوده است. یک مثال کلاسیک، هواشناسی است، که در آن مقدار بسیار عظیمی از داده‌های عددی باید بررسی گردند تا پیش‌بینی قابل اطمینانی از آب و هوا حاصل شود. به همین ترتیب مجموعه‌های مه‌داده در مدل‌های آب و هوا، ژئوفیزیک و نجوم ظاهر می‌شوند.

با این حال، مجموعه داده‌ها در این مسائل، هرچند که بسیار بزرگ هستند ولی دارای ساختار مناسب و قابل فهم، با سطوح معینی از عدم اطمینان، می‌باشند. دلیل این امر این است که آنها از فرآیندهای فیزیکی بدست آمده‌اند که در کل برای دانشمندان به خوبی فهمیده شده‌اند. چالش واقعی درک و برخورد با مه‌داده از علوم زیستی، علوم اجتماعی و به ویژه فعالیت‌های مبتنی بر مردم، پیش می‌آید. چنین داده‌هایی اغلب درهم، ناقص، غیر قابل اعتماد، پیچیده، تبدیل یافته بوده و به سرعت تولید می‌شوند. این داده‌ها اغلب کیفی نه کمی و ناهمگن بوده و در مورد روابط بین اشیاء به جای خود اشیاء، می‌باشند در حالی که داده‌های فیزیکی چنین نیستند.

^۶Smart watch ^۷uncertainties ^۸network theory ^۹node ^{۱۰}edges ^{۱۱}Twitter

زیرزمینی و روزمینی) و شبکه‌های زیست محیطی (زنجیره‌های غذایی، مکانیسم‌های بیماری‌ها و عفونت).

قدرت نظریه شبکه

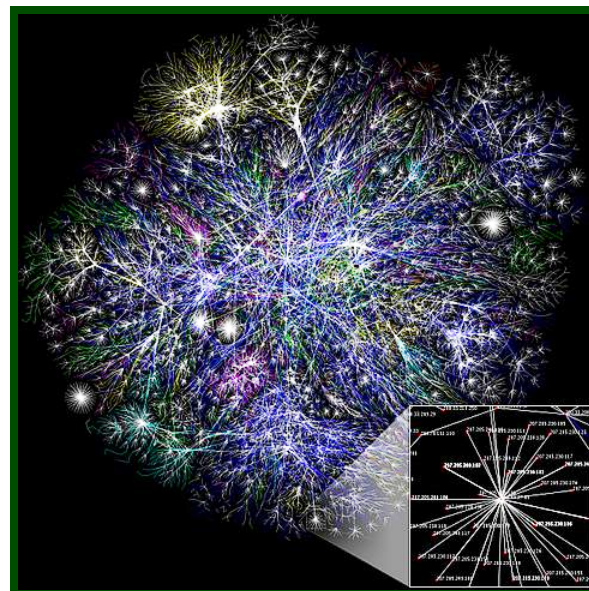
نظریه شبکه می‌تواند به بسیاری از دیگر سؤالات مرتبط با مه‌داده‌ها پاسخ دهد. هنگامی که شما با شبکه‌های بسیار بزرگ مواجه می‌شوید شناسایی خوشه‌ها^{۱۳} و طبقات — گروهی از گره‌ها که بسیار با هم در ارتباطند — یا تقسیم داده به گروه‌هایی که ویژگی‌های مشترکی داشته‌باشند، همیشه آسان نیست. چنین اطلاعاتی در داده کاوی^{۱۴} و شناخت الگوها^{۱۵} به خصوص در ارتباط با صنعت خرده‌فروشی، که به رفتار و ترجیحات مشتریان خود علاقه‌مندند، حیاتی است، اما همچنین می‌تواند برای شناسایی گروه‌های دوستی در شبکه‌های اجتماعی، تحقیق در سازمان مغز، و حتی در پیدا کردن الگوهای رای‌دهی در مسابقه آواز یورو ویژن مورد استفاده قرار گیرند. نظریه شبکه الگوریتم‌هایی برای شناسایی خوشه‌ها و برای طبقه‌بندی اطلاعات ارائه می‌کند.

علاوه بر این، چنین تجزیه و تحلیلی می‌تواند به مسأله بسیار مهم دیگری که در بسیاری از کاربردها با آن مواجه می‌شود کمک کند: پیوند دادن بانک‌های اطلاعاتی با سطوح مختلف دسته‌بندی داده‌ها در فضا و زمان. مثالی از این، پیش بینی وضعیت آب و هوا است، که در آن برخی از داده‌های عظیم در هر ثانیه توسط ماهواره‌های مدار زمین ارسال و انتقال می‌یابد. داده‌های دیگر ممکن است از ایستگاه‌های زمینی باشد که تنها چند عدد را اندازه‌گیری می‌کنند. برخی دیگر از داده‌ها حتی ممکن است قدمتی تاریخی داشته باشند، مانند سوابق کاپیتان‌ها در دریا در ۱۰۰ سال گذشته. تمام این سه مجموعه داده‌ها مفیدند و آنها را باید بطور یکپارچه باهم پیوند داد.



شکل ۴: اطلاعات، شایعات، بیماری‌های عفونی: همه آنها از طریق شبکه‌های اجتماعی گسترش می‌یابند.

تماس ارتباطی باشد. نظریه شبکه ماهیت شبکه را تشریح می‌کند و به ما اجازه می‌دهد تا به جستجوی ارتباط بین نقاط منفرد در مجموعه داده‌ها بپردازیم، و می‌تواند حرکت اطلاعات در گرداگرد شبکه را توصیف کند. (شما می‌توانید اطلاعات بیشتر در مورد گراف‌ها و شبکه‌ها را در وب سایت پلاس بخوانید.)



شکل ۳: نقشه اینترنت بر اساس داده‌های سال ۲۰۰۵ در پروژه CC BY-۵.۲ The Opte Project.

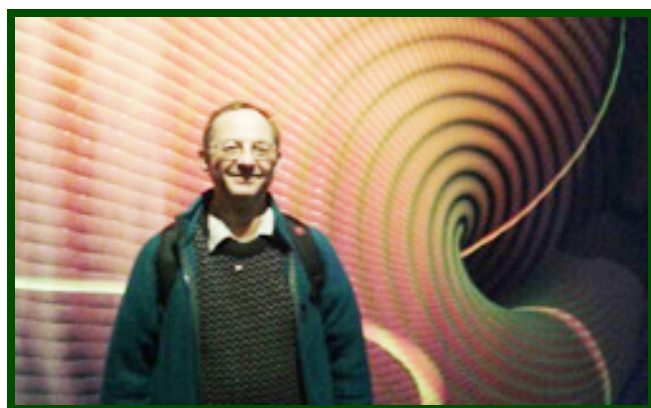
در واقع، اداره کردن شبکه تلفن همراه (که البته می‌تواند برای داندلود داده نیز مورد استفاده قرارگیرد) کاربرد قابل توجه و رو به رشد نظریه‌ی رنگ‌آمیزی گراف^{۱۲} است: پیدا کردن روش‌های رنگ-آمیزی یال‌ها یا گره‌ها در شبکه با توجه به قیود خاص، همچون گره‌های مجاور باید به رنگ‌های متفاوت باشند، صورت می‌گیرد. به عنوان مثال، رنگ ممکن است نشان دهنده فرکانس اختصاص داده‌شده به فرستنده‌های تلفن همراه باشد که باید طوری انتخاب شود تا تداخل را به حداقل رساند و بنابراین برای فرستنده‌های مجاور باید متفاوت باشد. رنگ‌آمیزی گراف تا همین اواخر در حوزه ریاضیات محض در نظر گرفته می‌شد.

مثال‌های دیگر از شبکه‌هایی که منجر به مه‌داده می‌شوند عبارتند از شبکه‌های سازمانی (مانند شبکه‌های مدیریت، جرایم سازمان یافته، حتی رفتار رأی‌دهی در مسابقات خوانندگی یورو ویژن)، شبکه‌های فن‌آوری (مانند شبکه برق یا مدارهای الکترونیکی)، شبکه‌های اطلاعاتی (متشکل از ژن‌ها، تعامل پروتئین با پروتئین، اشاعه کلمات مصطلح اطلاعات، افسانه‌ها و شایعات)، شبکه‌های حمل و نقل (مانند شرکت‌های هواپیمایی، تدارکات غذایی، سیستم‌های ریلی

¹²theory of graph coloring ¹³clusters ¹⁴data mining ¹⁵pattern recognition

استفاده قرار گیرد تا بدانیم چگونه اجزای مختلف برای تشکیل یک تصویر مناسب با همدیگر ترکیب می‌شوند. در زمینه یادگیری ماشین^{۲۱} این به ماشین اجازه می‌دهد تا «درک» کند که اشیاء در تصویر چیستند و تصمیمی «مستدل» درباره آن بگیرد.

این فقط لیستی کوتاه از کاربردها است. بسیاری از شاخه‌های دیگر علوم ریاضی و کامپیوتر وجود دارند که کاربردهایی در مه‌داده پیدا کرده‌اند. من مطمئنم که شاهد پیشرفت‌های بزرگی در ریاضیات محض، کاربردی و محاسباتی که ناشی از این چالش‌ها هستند خواهیم بود.



در مورد نویسنده

کریس باد^{۲۲} استاد ریاضیات کاربردی در دانشگاه باث^{۲۳}، رئیس مؤسسه ریاضیات و کاربردهای آن، رئیس ریاضیات مؤسسه سلطنتی و همکار افتخاری انجمن علمی بریتانیا است. به ویژه او علاقه‌مند به کاربرد ریاضیات در دنیای واقعی و ترویج درک عمومی از ریاضیات است. او نویسنده همکار کتاب معروف «تا دلت بخواد ریاضیات^{۲۴}» است که توسط انتشارات دانشگاه آکسفورد منتشر گردیده است.

* گروه آمار، دانشگاه زنجان

** دانشگاه خواجه نصیرالدین طوسی

سوالی به همان درجه اهمیت این است که چگونه یک شبکه به کل شبکه متصل شده است: آیا هر گره به بسیاری از گره‌های دیگر در سراسر شبکه متصل می‌شود یا ارتباطات پراکنده و تنک^{۱۶} هستند؟ کوتاه‌ترین مسیر در شبکه چیست؟ این سؤالات برای مسیریابی کارا در اینترنت، تفسیر داده‌های تدارکاتی، فهم سرعت ارتباطات و حتی بازاریابی، ضروری است. همچنین نقش نظریه شبکه برای جستجوی گره‌های با نفوذ در شبکه‌های بزرگ اساسی است. گره‌های بسیار مرتبط - چه آنها که نشان دهنده مردم، وب سایت‌ها هستند یا آنها که نشان دهنده فرودگاه‌ها هستند - برای استحکام شبکه بسیار مهم هستند چرا که حذف آنها ارتباطات کلی را بطور قابل توجهی تغییر می‌دهد. چنین اطلاعاتی می‌تواند به شکست سازمان‌های تروریستی، جلوگیری از شیوع و گسترش بیماری، یا حفظ ترافیک هوایی هنگامی که منطقه متاثر از شرایط بد آب و هوایی است بیانجامد.

ریاضیات چه چیز دیگری می‌تواند انجام دهد؟

نظریه شبکه فقط یکی از انواع فراوان تکنیک‌های ریاضی است که برای مطالعه مه‌داده استفاده می‌شود. بیشتر مه‌داده‌ها به فرمت تصویر هستند، بنابراین الگوریتم‌های ریاضی که برای طبقه بندی، تفسیر، تجزیه و تحلیل و فشرده سازی تصاویر مورد استفاده قرار می‌گیرند بسیار مهم هستند. روش‌های آماری^{۱۷} به مدت طولانی در تجزیه و تحلیل و تفسیر تصاویر استفاده شده‌اند اما اخیراً رشد قابل توجهی در الگوریتم‌های جدید ریاضی صورت گرفته است در حالی که این ایده‌ها از ریاضیات محض نشأت گرفته‌اند و مردم قبلاً فکر می‌کردند ریاضیات هیچ کاربرد مستقیمی در دنیای واقعی ندارد.

برخی از این الگوریتم‌ها براساس تحلیل معادلات پیچیده‌ای است که منجر به برخی کاربردهای قدرتمند و غیر منتظره از ابزارهای بسیار تکنیکی مربوط به نظریه معادلات^{۱۸} می‌شود. توپولوژی جبری^{۱۹}، شاخه‌ای از ریاضیات که به کمک جبر به بررسی خواص اشکال می‌پردازد، نقش بسیار مفیدی در طبقه‌بندی تصاویر ایفا می‌کند. نظریه رسته‌ها^{۲۰}، شاخه‌ای از ریاضیات که به بررسی ساختارها و مفاهیم ریاضی در سطح بسیار انتزاعی می‌پردازد، می‌تواند مورد