Normung

Titusz Pan*

ISCC: Neue Perspektiven für die KI-gesteuerte Identifikation von Inhalten

https://doi.org/10.1515/iwp-2024-2032

Zusammenfassung: Der International Standard Content Code (ISCC), der kürzlich als ISO 24138:2024 standardisiert wurde, stellt einen neuen Ansatz zur Identifizierung von Inhalten dar, indem er Identifikatoren direkt aus digitalen Inhalten generiert. Dieser Artikel befasst sich mit der Verwendung von tiefen neuronalen Netzen zur Erstellung von Semantischen Codes für Text (ISCC-SCT), die eine sprachübergreifende Identifizierung abstrakter Konzepte ermöglichen. Durch die Schließung der semantischen Lücke zwischen Daten und Informationen verbessert ISCC-SCT die Wissensorganisation und -abfrage. Der Artikel untersucht die Prinzipien von ISCC, den technischen Prozess der Generierung von semantischen Codes und seine möglichen Anwendungen in der Bibliotheks- und Informationswissenschaft (LIS).

Deskriptoren: Norm, Neuronales Netz, Kennwert, Text, Daten, Information,Künstliche Intelligenz, Wissensmanagement, Simprint

ISCC: New perspectives for the AI-driven identification of content

Abstract: The International Standard Content Code (ISCC), recently standardized as ISO 24138:2024, introduces a novel approach to content identification by generating identifiers directly from digital content. This article focuses on the ISCC's use of deep neural networks to create Semantic Codes for Text (ISCC-SCT), enabling the cross-lingual identification of abstract concepts. By addressing the semantic gap between data and information, the ISCC-SCT enhances knowledge organization and retrieval. The article explores the ISCC's principles, the technical process of Semantic Code generation, and its potential applications in Library and Information Science (LIS).

Descriptors: Standard, Neural network, Identifier, Text, Data, Information, Artificial intelligence, Knowledge management, SimprintISCC : nouvelles perspectives pour l'identification de contenu guidée par l'IA

***Kontaktperson: Titusz Pan**, CRAFT AG, Wallstr. 6, 79098 Freiburg, E-Mail: tp@craft.de. https://orcid.org/0000-0002-0521-4214

Résumé : Le Code international normalisé du contenu (ISCC), récemment normalisé sous la forme de la norme ISO 24138:2024, introduit une nouvelle approche de l'identification du contenu en générant des identifiants directement à partir du contenu numérique. Cet article se concentre sur l'utilisation par l'ISCC de réseaux neuronaux profonds pour créer des codes sémantiques pour le texte (ISCC-SCT), permettant l'identification interlinguistique de concepts abstraits. En comblant le fossé sémantique entre les données et les informations, l'ISCC-SCT améliore l'organisation et la récupération des connaissances. L'article explore les principes de l'ISCC, le processus technique de génération de codes sémantiques et ses applications potentielles en bibliothéconomie et sciences de l'information (LIS).

Descripteurs : Norme, Réseau neuronal, Identificateur, Texte, Données, Information,Intelligence artificielle, Gestion des connaissances, Simprint

Die Daten- und Wissens Krise

Das exponentielle Wachstum von Daten und die Informationsüberflutung

Im digitalen Zeitalter hat das Wachstum der Datenmenge ein beispielloses Ausmaß erreicht. Das globale Datenvolumen wächst exponentiell, und Schätzungen zufolge wird die Welt bis zum Jahr 2025 täglich 463 Exabyte an Daten erzeugen[1]. Dieses explosionsartige Datenwachstum, das durch die zunehmende Digitalisierung nahezu aller Lebensbereiche vorangetrieben wird, stellt sowohl Einzelpersonen als auch Organisationen und die Gesellschaft insgesamt vor erhebliche Herausforderungen. Das Phänomen der Informationsüberflutung, bei dem das schiere Volumen der verfügbaren Daten unsere Fähigkeit, diese zu verarbeiten und zu verstehen, übersteigt, ist zu einem prägenden Merkmal der modernen Ära geworden.

1 Desjardins, J. (2019). How Much Data Is Generated Each Day? World Economic Forum. https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/ [28.8.2024].

Das Paradox des Wissensmanagements

Trotz der enormen Vielfalt an Daten gibt es ein Paradoxon, das unsere Fähigkeit behindert, diese Daten effektiv in Wissen umzuwandeln. Die riesigen Datenmengen aus unterschiedlichsten Quellen sollten theoretisch die Entscheidungsfindung und Innovation fördern. Der überwältigende Umfang und die Komplexität dieser Datenflut behindern jedoch häufig die Gewinnung sinnvoller Erkenntnisse. Herkömmliche Methoden der Datenorganisation, -klassifizierung und -abfrage geraten zunehmend an ihre Grenzen, so dass eine Lücke zwischen dem Potenzial der verfügbaren Daten und ihrem praktischen Nutzen für das Wissensmanagement entsteht.

Die Notwendigkeit für neue Ansätze

Traditionelle Identifier wie ISBN, DOI[2] und andere spielen eine zentrale Rolle im Management und der Organisation von Informationen. Diese Identifikatoren haben eine standardisierte, verlässliche und interoperable Methode zur Katalogisierung und Referenzierung von Werken über eine Vielzahl von Disziplinen hinweg bereitgestellt. Sie bilden die Grundlage für unzählige Systeme in Bibliotheken, Archiven und digitalen Plattformen und sind somit für die Bibliotheks- und Informationswissenschaft (LIS) unverzichtbar. Allerdings operieren diese Identifikatoren in erster Linie auf der Ebene abstrakter Konzepte – sie identifizieren Bücher, Artikel und andere intellektuelle Werke und werden manuell kuratiert. Seit der Einführung dieser Identifikationssysteme ist viel Zeit vergangen und die Landschaft hat sich erheblich verändert, was eine deutliche Lücke offenbart: die Fähigkeit, die eigentlichen digitalen Daten, die unsere Inhalte codieren – jene Dateien, Bilder, Videos und Artikel, die das Internet und die unzähligen institutionellen Datensammlungen bevölkern – effizient zu identifizieren und zu verwalten.

Wir benötigen neue Systeme, die dort ansetzen und ergänzen, wo traditionelle Identifikatoren an ihre Grenzen stoßen – Systeme, die auf der Ebene der Daten selbst operieren. Hier kommt die nächste Generation von Identifikatoren ins Spiel, die durch Fortschritte in der Datenwissenschaft und Künstlichen Intelligenz angetrieben wird.

Die semantische Kluft

Daten, Information und der Interpreter

Um die semantische Kluft[3] besser zu verstehen, möchte ich zunächst einige sehr spezifische Definitionen der Begriffe Daten, Information und Interpreter im Kontext dieses Artikels einführen:

Daten[4]: Rohe, unverarbeitete Sequenzen von Symbolen, die sowohl in digitaler als auch in analoger Form existieren können. Daten existieren in einer physischen und messbaren Form.

Information: Die aus Daten abgeleitete Bedeutung, die abstrakte und immaterielle Konzepte wie Ideen, Wissen und Verständnis repräsentiert. Information ist von Natur aus kontextabhängig.

Interpreter: Die Entität, sei es Mensch oder Maschine, welche Information durch Anwendung spezifischer Regeln, Formate oder Algorithmen codiert oder decodiert, um Sequenzen von Symbolen zu erzeugen oder zu konsumieren – Symbole, die gespeichert, übertragen oder verarbeitet werden können. Der Interpret fungiert als Brücke zwischen Daten und Information, wobei Menschen auf kognitive Fähigkeiten und Maschinen auf Algorithmen und Datenverarbeitung angewiesen sind.

Diese Trinität ist entscheidend, um die Herausforderungen der Identifikation von Inhalten zu verstehen. Während Daten konkret und messbar sind, ist Information abstrakt und kontextabhängig. Die semantische Lücke besteht zwischen Daten und Information, wobei der Interpret die entscheidende Brücke darstellt, um Bedeutung zu erschließen.

Die menschliche Sichtweise

Der Mensch ist von Natur aus eher an Informationen als an Rohdaten interessiert. Unsere kognitiven Prozesse sind darauf ausgerichtet, Bedeutungen zu extrahieren, Muster zu erkennen und abstrakte Konzepte zu bilden. Diese Vorliebe für Abstraktion ist tief in unserer Evolutionsgeschichte und kognitiven Architektur verwurzelt.

Wenn wir zum Beispiel ein Buch lesen, konzentrieren wir uns nicht auf die einzelnen Buchstaben oder Wörter

2 Paskin, N. (2010). Digital Object Identifier (DOI®) System. Encyclopedia of Library and Information Sciences, 3, 1586–1592

3 Hein, Andreas M. „Identification and bridging of semantic gaps in the context of multi-domain engineering.“ Forum on Philosophy, Engineering & Technology. 2010.

4 Ackoff, Russell L. „From data to wisdom.“ Journal of applied systems analysis 16.1 (1989): 3–9.

(die Daten), sondern auf die Geschichte, die Ideen oder Argumente, die präsentiert werden (die Informationen). Diese Fähigkeit, zu abstrahieren und zu konzeptualisieren, ermöglicht es uns, komplexe Ideen zu verarbeiten und neues Wissen zu erlangen.

Dieser Umstand stellt uns im digitalen Zeitalter jedoch auch vor Herausforderungen. Als Informations-Experten müssen wir Systeme entwickeln, die die Kluft zwischen den Rohdaten, welche durch Maschinen verarbeitet werden, und den abstrakten Informationen, die Menschen interessieren, überbrücken können.

Herausforderungen bei der Identifikation von Information

Eine der größten Herausforderungen für die Identifizierung und Verwaltung von Inhalten liegt in der semantischen Kluft, d. h. der Diskrepanz zwischen den Rohdaten und den konzeptionellen Informationen, die sie darstellen:

Kontextuelle Variabilität: Ein und dasselbe Datenelement kann je nach Kontext und Interpreter unterschiedliche Informationen liefern. So kann beispielsweise die Zahl „42“ nur eine zufällige ganze Zahl sein, ein Verweis auf ein Werk von Douglas Adams[5] oder eine kritische Messung in einer wissenschaftlichen Arbeit.

Sprachliche und kulturelle Unterschiede: Informationen sind oft sprach- und kulturabhängig, was das sprach- und kulturübergreifende Informationsmanagement zu einer komplexen Aufgabe macht.

Sich wandelnde Bedeutungen: Die Interpretation von Informationen kann sich im Laufe der Zeit ändern, wenn sich die gesellschaftlichen Normen, das wissenschaftliche Verständnis oder die kulturellen Bezüge weiterentwickeln.

Implizites Wissen: Ein Großteil des menschlichen Verständnisses beruht auf implizitem Wissen oder „gesundem Menschenverstand“, das sich nur schwer kodifizieren und in Informationssysteme einbinden lässt.

Diese Herausforderungen unterstreichen den Bedarf an hochentwickelten Systemen zur Identifizierung von Inhalten, welche die nuancierte, kontextabhängige Natur von Information erfassen und darstellen können.

Die semantische Lücke in digitalen Systemen

Herkömmliche digitale Systeme sind hervorragend in der Lage, Daten zu speichern und zu verarbeiten, haben aber Schwierigkeiten, den semantischen Inhalt dieser Daten abzubilden. Diese Diskrepanz zwischen der maschinellen Darstellung und dem menschlichen Verständnis zeigt sich auf verschiedene Weise:

Stichwort basierte Suche: Herkömmliche Suchmaschinen stützen sich oft auf den Abgleich von Schlüsselwörtern und übersehen dabei die kontextuellen Nuancen, die Menschen von Natur aus verstehen.

Unzulängliche Metadaten: Manuell erstellte Metadaten reichen oft nicht aus, um den gesamten semantischen Reichtum von Inhalten zu erfassen, insbesondere bei multimedialen Daten.

Format Inkompatibilitäten: Unterschiedliche Datenmodalitäten (Text, Bild, Audio, Video) werden oft getrennt behandelt, so dass potenzielle semantische Verbindungen zwischen den Formaten fehlen.

Skalierbarkeit: Mit wachsendem Datenvolumen werden manuelle Methoden zum Erfassen von semantischem Kontext zunehmend unpraktisch.

Auf dem Weg zu maschinellem Abstraktionsvermögen

So wie die kryptischen Symbole einer fremden Schrift für Unkundige keine Bedeutung haben, sind digitale Bits auf einem Speichermedium für Maschinen ohne die Programme, die sie interpretieren, bedeutungslos.

Ähnlich wie die Evolution die menschliche Intelligenz im Laufe der Jahrtausende geformt hat, werden moderne neuronale Netze durch riesige Datenmengen geformt. Diese Systeme sind zunehmend in der Lage, Abstraktionen in Daten zu erkennen und darzustellen.

Die bemerkenswerten Fortschritte der Künstlichen Intelligenz (KI), insbesondere bei großen Sprachmodellen (LLMs), hat viele dazu veranlasst, das Potenzial von Maschinen neu zu bewerten. Es gilt nicht mehr als unvorstellbar, dass diese Systeme irgendwann menschliche Intelligenz erreichen und vielleicht sogar eines Tages übertreffen.

Das grundlegende Trainingsziel dieser Sprachmodelle – die Vorhersage des nächsten Tokens für eine gegebene Texteingabe – mag täuschend einfach erscheinen, was manche Leute sogar dazu veranlasst, sie als „statistische Papageien“ zu bezeichnen.

Um jedoch das jeweils nächste Token überzeugend vorhersagen zu können, muss das neuronale Netz verborgene Abstraktionen in den zugrundeliegenden Daten erkennen

5 In Douglas Adams Science-Ficton Klassiker „Per Anhalter durch die Galaxis“ ist die Zahl 42 die von einem Supercomputer errechnete Antwort auf die „endgültige Frage nach dem Leben, dem Universum und dem ganzen Rest“: Douglas Adams: Per Anhalter durch die Galaxis. Heyne, München 2009. ISBN 978-3-453-14697-6.

und zuordnen – ein Prozess, der weitaus komplexer ist, als es zunächst scheint.

An dieser Stelle wird es für Informations-Experten spannend. Neuronale Netze transformieren Daten, indem sie verborgene Abstraktionen intern in Form von sog. „latenten Repräsentationen" materialisieren[6]. Und genau diese „latenten Darstellungen" können als leistungsfähiges neues Instrument zur Überbrückung der semantischen Kluft dienen.

Die Herausforderung für Informations-Experten wird darin bestehen, diese KI-Technologien effektiv zu nutzen, um besser durch die immer komplexere Wissenslandschaft der Menschheit zu navigieren. Die Zukunft liegt in Lösungen, die nahtlos zwischen der Welt der Daten und dem menschlichen Verstehen vermitteln können und so ein intuitiveres und effektiveres Informations-Ökosystem erschaffen.

Einführung zu Simprints

Um den komplexen Herausforderungen der Identifizierung digitaler Inhalte – wie dem effizienten Auffinden und Vergleichen großer Datenmengen – zu begegnen, führen wir einen neuen Begriff ein: **Simprint**.

Simprints repräsentieren eine Familie von Algorithmen, die Daten in kompakte binäre Codes komprimieren, die eine spezifische Ähnlichkeits-Struktur der Daten beibehalten. Diese digitalen Signaturen identifizieren nicht nur Inhalte, sondern fassen auch deren wesentliche Merkmale zusammen, wodurch ein effizienter Vergleich und eine einfache Wiederauffindbarkeit ermöglicht werden.

Obwohl „Simprint" in diesem Artikel als neuer Begriff eingeführt wird, werden die zugrundeliegenden Technologien schon seit langem in verschiedenen wissenschaftlichen Bereichen verwendet, oft unter verschiedenen Namen wie Simhash[7], Perceptual Hash, Locality-sensitive Hashes, Binary Embeddings[8] und weitere.

Diese Konzepte tauchen in verschiedenen Disziplinen auf, was zu einem Mangel an einheitlicher Terminologie führt, was wiederum die disziplinübergreifende Kommunikation und die Entwicklung integrierter Lösungen behindert.

Wir prägen den Begriff „Simprint" – eine Mischung aus „Similarity" und „Fingerprint" – als einheitliches Konzept für diese Familie von Technologien. Unser Ziel ist es, einen klaren, intuitiven Begriff zu schaffen, der die Kommunikation und das Verständnis innerhalb der Bibliotheks- und Informationswissenschaft und darüber hinaus fördert und diese leistungsstarken Werkzeuge für die Identifikation von Inhalten einem breiteren Publikum zugänglich macht.

Es ist wichtig, Simprints nicht mit kryptografischen Hash-Funktionen wie MD5 oder SHA-256 zu verwechseln, die selbst bei geringsten Variationen in den Eingabedaten völlig unterschiedliche Ausgaben erzeugen.

Der wesentliche Vorteil von Simprints liegt in ihrer Fähigkeit, Ähnlichkeitsbeziehungen zwischen den ursprünglichen Eingaben auf verschiedenen Abstraktionsebenen zu materialisieren. Dies macht sie besonders leistungsfähig für inhaltsbasierte Indexierung und Suche, insbesondere beim Management von groß angelegten digitalen Bibliotheken und Archiven. Durch die Einführung von Simprints kann der Prozess der Identifikation von Inhalten optimiert werden.

Die vielschichtige Natur von „Ähnlichkeit"

Ähnlichkeit ist, ähnlich wie Schönheit, subjektiv und hängt von der Perspektive des Interpreters ab. Im Zusammenhang mit digitalen Inhalten kann Ähnlichkeit durch mindestens drei miteinander verbundene Dimensionen verstanden werden: Daten-Ähnlichkeit, syntaktische Ähnlichkeit und semantische Ähnlichkeit.

Daten-Ähnlichkeit: Auf der grundlegendsten Ebene umfasst die Daten-Ähnlichkeit den Vergleich der rohen, unverarbeiteten Bitströme digitaler Dateien. Statistische Algorithmen bewerten die Ähnlichkeit, indem sie die Sequenzen von Bits und Bytes analysieren, ohne auf die Bedeutung des Inhalts einzugehen. Daten-Ähnlichkeit ist unerlässlich für die Identifizierung von Duplikaten und die Gewährleistung der Datenintegrität.

Syntaktische Ähnlichkeit: Auf einer höheren Ebene untersucht die syntaktische Ähnlichkeit die wahrnehmbaren, strukturellen und syntaktischen Merkmale digitaler Medien nach Dekodierung der Rohdaten. Dies kann den Vergleich von lexikalischen Mustern in Texten, visuellen Elementen in Bildern oder auditiven Merkmalen in Audiodateien umfassen, wobei oft sorgfältig manuell konzipierte Algorithmen zum Einsatz kommen.

Semantische Ähnlichkeit: Auf der abstraktesten Ebene geht die semantische Ähnlichkeit über die wahrnehmbaren und strukturellen Aspekte hinaus und berücksichtigt die Bedeutung und den Kontext des Inhalts. Diese

6 Huang, Yuzhen, et al. „Compression represents intelligence linearly." arXiv preprint arXiv:2404.09937 (2024).

7 Sadowski, Caitlin, and Greg Levin. „Simhash: Hash-based similarity detection." (2007).

8 Kusupati, Aditya et al. „Matryoshka Representation Learning." Neural Information Processing Systems (2022).

Art der Ähnlichkeit konzentriert sich auf hochrangige Konzepte und Ideen, die in digitalen Medien dargestellt werden, und verwendet fortschrittliche Deep-Learning-Techniken, um den Inhalt zu analysieren und zusammenzufassen.

Innerhalb jeder dieser Kategorien können verschiedene Simprint-Algorithmen entwickelt werden, die jeweils einen unterschiedlichen analytischen Ansatz auf die Eingabedaten anwenden und diese nach bestimmten Ähnlichkeitsdimensionen clustern.

Die Ähnlichkeit kann sowohl global als auch partiell bewertet werden. Die globale Ähnlichkeit bewertet die allgemeine Ähnlichkeit zwischen zwei Dateien in ihrer Gesamtheit, während die partielle Ähnlichkeit sich auf bestimmte Segmente oder Regionen innerhalb des Inhalts konzentriert. Die globale Ähnlichkeit kann beispielsweise zum Vergleich ganzer Dokumente in einer digitalen Bibliothek verwendet werden, während die partielle Ähnlichkeit für die Identifizierung ähnlicher Passagen oder Abschnitte in verschiedenen Texten entscheidend sein kann.

Diese Flexibilität bei der Ähnlichkeits-Bewertung ermöglicht es uns, unseren Ansatz auf die spezifischen Anforderungen unserer Sammlungen oder Anwendungsfälle abzustimmen. In einer digitalen Bibliothek mit historischen Dokumenten könnte beispielsweise inhaltliche Ähnlichkeit verwendet werden, um visuell ähnliche Manuskripte zu finden, Daten-Ähnlichkeit, um exakte digitale Kopien zu identifizieren, und semantische Ähnlichkeit, um thematisch verwandte Texte in verschiedenen Sprachen oder Zeiträumen zu finden.

Durch die Nutzung dieser verschiedenen Arten von Ähnlichkeit bieten Simprints einen leistungsstarken, nuancierten Ansatz für die Organisation, den Vergleich und die Suche nach digitalen Inhalten und eröffnen neue Möglichkeiten für die Verwaltung und Erkundung unserer ständig wachsenden digitalen Sammlungen.

Simprints als Identifikatoren

Es ist wichtig zu erkennen, dass Simprints ein bestimmtes digitales Artefakt nicht eindeutig identifizieren. Sie zeichnen sich dadurch aus, dass sie Informationen auf einer höheren Abstraktionsebene identifizieren, anstatt nur Rohdaten. Ein einzelner Simprint fasst mehrere ähnliche Inhalte unter einer gemeinsamen Kennung zusammen. Diese Eigenschaft verleiht Simprints ihre Stärke bei der Organisation und dem Auffinden von Informationen basierend auf Ähnlichkeiten.

Beispielsweise würden in einem digitalen Bildarchiv mehrere Bilder mit ähnlichen visuellen Merkmalen – wie etwa verschiedene Fotos desselben Objekts – unter einem einzigen Simprint gruppiert, was eine effiziente Wiederauffindung und den Vergleich erleichtert.

Wenn jedoch ein Simprint mit einem kryptografischen Hash kombiniert wird – ein Algorithmus, der eine eindeutige, festgelegte Zeichenfolge aus beliebigen Eingangsdaten generiert – entsteht ein leistungsstarker Hybrid-Identifikator. Dieser Hybrid vereint die Fähigkeit der eindeutigen Identifizierung von Daten mit der Ähnlichkeits-basierten Gruppierung in einer einzigen, kompakten Zeichen folge.

In einer digitalen Bibliothek könnte dieser Hybrid-Identifikator beispielsweise sowohl die präzise Auffindung eines bestimmten Dokuments (dank des kryptografischen Hashes) als auch die Möglichkeit bieten, verwandte Dokumente zu erkunden, die thematische oder strukturelle Ähnlichkeiten aufweisen (ermöglicht durch den Simprint). Diese duale Funktionalität verbessert sowohl die Genauigkeit als auch die Suchmöglichkeiten im digitalen Inhaltsmanagement.

Vorstellung des ISCC

Der International Standard Content Code (ISCC – ISO 24138:2024) ist ein System zur Identifizierung von Inhalten. Er basiert auf einer Sammlung von vordefinierten Simprints mit einer gemeinsamen Syntax und einer selbst beschreibenden Struktur.

Das Konzept des ISCC entstand 2016 im Rahmen eines journalismusbezogenen Forschungsprojekts. In Anerkennung seines Potenzials schlug die deutsche Normungsorganisation DIN der Internationalen Organisation für Normung (ISO) eine Initiative zur weiteren Erforschung dieser Idee vor. Die formale Arbeit an der Norm begann im Oktober 2019 im ISO/TC 46/SC 9 Information und Dokumentation, dem gleichen Ausschuss, der auch für weit verbreitete Standards wie die ISBN und DOI verantwortlich ist. Nach mehr als vier Jahren gemeinsamer Arbeit wurde ISO 24138:2024 im Mai 2024 veröffentlicht. Dies ist ein wichtiger Meilenstein, der zum ersten Mal eine standardisierte Methode zur effizienten und interoperablen Identifizierung von digitalen Inhalten in verschiedenen Sektoren ermöglicht.

Der ISCC ist auf digitale Artefakte jeder Granularität anwendbar und wird direkt aus den Daten mithilfe einer Reihe von Open-Source-Algorithmen erstellt. Im Gegensatz zu einem monolithischen Identifikator ist der ISCC ein hybrider Code, der sich aus mehreren Einheiten zusammensetzt, von denen jede eine bestimmte Aufgabe bei der Identifizierung von Inhalten erfüllt. Diese Einheiten, bekannt als ISCC-UNITS, arbeiten zusammen, um eine umfassende,

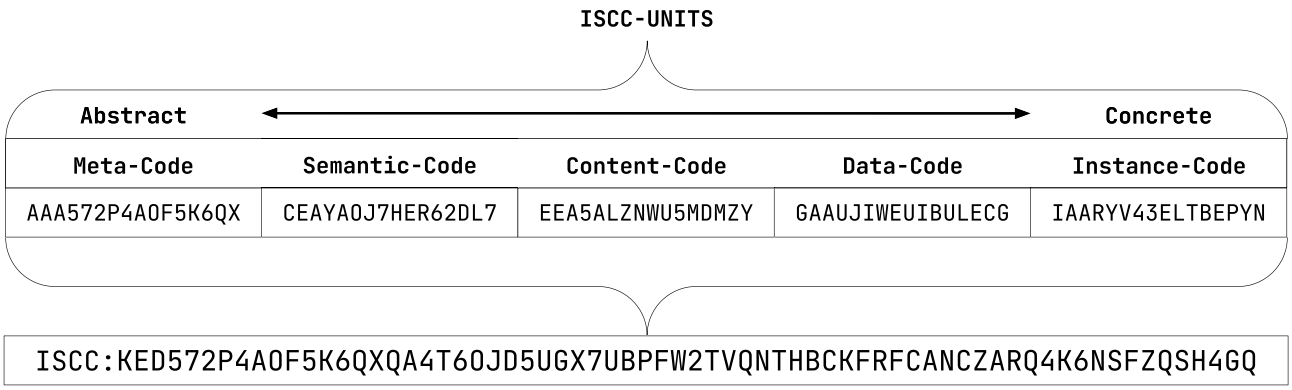

Abbildung 1

kompakte und differenzierte „Beschreibung“ digitaler Inhalte zu schaffen.

Ein wesentliches Merkmal des ISCC ist seine Unterstützung für Multimodalität, die spezialisierte Algorithmen für alle wichtigen Inhaltstypen umfasst: Text, Bild, Audio und Video. Diese multimodale Fähigkeit stellt sicher, dass der ISCC unabhängig von der Art des Inhalts auf ein breites Spektrum digitaler Artefakte angewendet werden kann.

Jedes Zeichen eines ISCC codiert 5 Bit an Informationen, was es ermöglicht, ISCCs für Ähnlichkeitsvergleiche des identifizierten Inhalts zu verwenden. ISCC-UNITS können unabhängig voneinander verwendet werden oder zu einem ISCC-CODE kombiniert werden, um ein breiteres Feld abzudecken.

ISCC-CODES werden ausschließlich zur Identifizierung digitaler Artefakte verwendet und sind nicht an einen bestimmten Herausgeber oder Eigentümer gebunden. Eine Registrierung ist nicht erforderlich, da der ISCC algorithmisch aus den Daten abgeleitet wird, was es jedem ermöglicht, ISCC-CODES für ein bestimmtes Artefakt zu erstellen und zu überprüfen. Unternehmen und Institutionen sind frei, ihre eigenen Register basierend auf ihren spezifischen Anforderungen zu entwickeln.

ISCC-SCT: Semantische Codes für Texte

Obwohl ISO 24138:2024 noch keine spezifischen Algorithmen für semantische ISCC-UNITs definiert, wird bereits eine Kategorie für „Semantische Codes“ reserviert. Die ISCC Foundation, eine gemeinnützige Organisation, die die Entwicklung des ISCC unterstützt, arbeitet im Rahmen ihrer ISCC-LAB-Projekte an semantischen ISCC-UNITs.

Eine wesentliche Innovation ist der mehrsprachige ISCC-SCT (Semantic-Code Text), der die Ähnlichkeit zwischen Texten in verschiedenen Sprachen erkennen kann und somit aussagekräftige Vergleiche über Sprachgrenzen hinweg ermöglicht.

Der ISCC-SCT Prototyp generiert kompakte Simprints, die über Sprachgrenzen hinweg robust, unempfindlich gegenüber Paraphrasierungen und in der Lage sind, Texte auf granularer Ebene für einen mehrsprachigen Informationsabruf zu verarbeiten. Dies macht ihn nützlich für das Indexieren, Suchen und Abgleichen von Texten, egal ob über Sprachgrenzen hinweg, paraphrasiert oder segmentiert.

Technischer Überblick: Vom Text zu semantischen Codes

Die Generierung eines semantischen Text-Codes umfasst mehrere Schritte:

Chunking: Der Eingabetext wird in Token, in der Regel Wörter oder Silben, zerlegt und zu Abschnitten von mehreren Sätzen zusammengefasst.

Embedding: Diese Abschnitte werden mithilfe eines neuronalen Netzwerks, das auf parallelen Sätzen in mehr als 50 Sprachen trainiert wurde[9], in einen hochdimensionalen Vektorraum eingebettet. Dadurch entstehen sprachunabhängige Repräsentationen, die die Semantik des Textes unabhängig von der Originalsprache erfassen.

Pooling: Die einzelnen Vektoren werden zu einem einzigen hochdimensionalen Vektor zusammengefasst, der als umfassende, sprachunabhängige Darstellung des semantischen Inhalts des gesamten Textes dient.

Binarization: Der globale Vektor wird in einen kompakten binären Code umgewandelt, um ihn für eine effiziente Speicherung und den Vergleich zu komprimieren, wobei wichtige semantische Informationen erhalten bleiben.

9 Reimers, Nils, and Iryna Gurevych. „Making monolingual sentence embeddings multilingual using knowledge distillation.“ arXiv preprint arXiv:2004.09813 (2020).

Der resultierende semantische Text-Code ist ein leistungsstarker Identifikator für semantische Informationen, der den Textvergleich und die -suche über Sprachen und Kontexte hinweg ermöglicht. Dies eröffnet neue Möglichkeiten für mehrsprachige Informationssysteme, in denen Inhalte ohne sprachspezifische Einschränkungen abgerufen und analysiert werden können.

Um das Engagement zu fördern, hat die ISCC Foundation eine interaktive Experimentierumgebung entwickelt, mit welcher Forscher und andere Interessierte diese Technologie erkunden können. Die Experimentierumgebung ist öffentlich verfügbar unter https://huggingface.co/spaces/iscc/iscc-sct und bietet einen praktischen Einstieg in die Generierung und den Vergleich von semantischen ISCC Text-Codes.

Herausforderung und Zukunft des ISCC

ISCC und die in Entwicklung befindlichen Semantic-Codes bieten spannende Möglichkeiten, aber es gibt auch Herausforderungen, die es zu berücksichtigen gilt:

Integration in bestehende Systeme: Die Einführung des ISCC erfordert die Integration in bestehende Systeme und digitale Repositorien, was sowohl technische als auch organisatorische Herausforderungen mit sich bringt.

Skalierbarkeit und Leistung: Da das Volumen digitaler Inhalte weiterhin exponentiell ansteigt, wird es entscheidend sein, die effiziente Erstellung und den Abgleich von ISCCs in großem Umfang zu gewährleisten.

Ethische Implikationen: Die Verwendung von KI-generierten Identifikatoren wirft Fragen zu Voreingenommenheit und Fairness auf. Es wird eine ständige Herausforderung sein, dafür zu sorgen, dass semantische Codes bestehende Verzerrungen in den Trainingsdaten nicht fortführen oder verstärken.

Fazit

Die Entwicklung des ISCC und der semantischen Codes stellt einen bedeutenden Fortschritt im Bereich der Identifikation von Inhalten dar. Diese Technologien bieten neue Methoden zur Verwaltung, Auffindung und zum Verständnis digitaler Informationen, mit dem Potenzial, die Effizienz und Genauigkeit dieser Prozesse zu verbessern.

Allerdings wird für die erfolgreiche Integration dieser Werkzeuge mehr als nur technologische Innovation erforderlich sein. Es bedarf eines kollaborativen Ansatzes, der sorgfältig das Gleichgewicht zwischen den Fähigkeiten der KI, menschlicher Expertise und ethischen Überlegungen abwägt.

Während sich diese Technologien weiterentwickeln, wird die Rolle der Bibliotheks- und Informationswissenschaft von entscheidender Bedeutung sein, um ihre Anwendung zu gestalten und sicherzustellen, dass sie effektiv zur Organisation und zum Zugang zu Wissen im digitalen Zeitalter genutzt werden.

Der Übergang von der Verwaltung großer Datenmengen hin zu einem semantischen Verständnis ist ein fortlaufender Prozess, der sowohl Beiträge künstlicher als auch menschlicher Intelligenz erfordert. Es ist entscheidend, diese Aufgabe mit Sorgfalt, Kreativität und einem verantwortungsvollen Einsatz dieser Werkzeuge anzugehen, um sicherzustellen, dass diese Fortschritte der Gesellschaft als Ganzes zugutezukommen.

Titusz Pan
ISCC Foundation
Demmersweg 92
NL 7556 BN Hengelo
tp@iscc.io
https://orcid.org/0000-0002-0521-4214

Titusz Pan ist Gründer und Vorstand der CRAFT AG mit Sitz in Freiburg und arbeitet seit dem Jahr 2000 an zahlreichen Medientechnologie-Projekten. Im Rahmen seiner Aktivitäten als Open-Source Programmierer hat er den ISCC entworfen und ist Gründer und Vorstand der ISCC Foundation.