Metamorph: Theory of Operation

Premise

In the days of a limited flow of printed information, a sharp observer could stay ahead through astute comprehension and good organization of materials. Managing a business, analyzing trends, or forecasting investments was all possible for the competent person staying on top of one's job.

But when one considers the sheer volume of text generated in a normal day of life in the nineties, one might hesitate before drawing a definitive conclusion about anything. There is hardly a profession which does not rely in some way on digitized information as a resource to stay current. This is true from law to medicine, from real estate to politics.

Certain text defies databasing: text which is nonpositional or otherwise resistive of fielding and categorization, text which is context rich and contains implicit knowledge. Not only is it difficult to create a structure for storing text which is as efficient as actual content when all possible interconnections are desired, but knowledge engineering hours cost dearly when the need is immediate.

With the advent of word processing, fax boards, scanners, OCR technology, and online data sources as automated solutions to information overload, electronic text is being used to lower costs and raise efficiency. But only when text can be accessed as freely as it was input in the first place can a dynamic information system be deemed truly useful.

Why Metamorph?

Metamorph is the best solution to text problems where meaning is an important part of the task. It contains built-in intelligence without knowledge engineering, and retains contextual richness which might be otherwise lost in a less resilient indexing system.

Combining techniques as appropriate, the Metamorph software package integrates a free text scan, hypertexting, and concept based technologies, optionally utilizing an inverted index only for optimization based on size of textbase.

Metamorph finds concepts, phrases, wildcarded strings, quantities as they exist in text, approximations (including typos, transpositions, and misspellings), variable length patterns, special expressions, and meaningful responses to natural language queries. Its user syntax is simple, while its search engine locates a multitude of patterns adequate to match the complexity of possibilities inherent in the English language.

Document Content

Documents are constructed from ideas. These ideas are communicated in sections made up of paragraphs, comprised of sentences created from a broad vocabulary. As language has evolved through centuries of civilization we have used words in context with other words in an ordered stream which we hope will transmit meaning to others.

Digitized narrative text records this transmission of thought, using every punctuation and space marking along with the exact word forms in the sequence they were written down to accurately represent the intended message. No matter how casually written, it is the ordered presentation of content which contains the maximum nuance and connotation of the author.

While various methods exist for efficiently storing lexical items out of context, the retrieval method is incomplete if it does not include an in context check for validity.

Metamorph makes use of the inherent organization already present in narrative text. Items are retrieved in association with other items within the smaller or greater unit the author chose to record his ideas or facts. Document organization such as line breaks, sentence endings, space between paragraphs, page formatting characters, and section headers which were already used to create the documents in the first place can be used to demarcate concept groupings within which associated search items lie.

In this sense, the formatting of the units of communication used to construct the document carry as much information as the words themselves. Additional tagging or preparation is not required if these natural markings can be used intelligently by the retrieval software.

When Metamorph searches a document it reads in a buffer of information, as it is, in sequence. It then looks for search items in proximity to each other inside of a natural delimiter such as a sentence or paragraph. Since all that information is present, all associations are possible, unlike an inverted file index which knows only isolated occurrences of strings in documents. Even where indexing is used at a first pass for search speed optimization, a final read of the text is always done to locate items in relation to their context.

The Quest for Relevance

A computer doesn't think; a person does. The task of retrieval software is to quickly provide information as close as possible to the question at hand, so that the human can read it, make associations, draw conclusions, and perhaps even come to a new realization.

To achieve this task, the program must present relevant choices. The more relevant the information presented, the better the software is deemed to have done its job. It is not the software's job to evaluate meaning, but rather to locate possible meaningful matches and present them for review.

So, where does this meaning lie, and how does the software find it? The answer comes back to the importance of context, the most efficient storage method if you wish to retain all possible connections for evaluation.

Words of themselves have several meanings. Look up any word in a dictionary exclusive of the way in which it was used, and it will be difficult to guess what it means. One can't expect a software program to know what even a human can't reason out. But plug the word into its context and the meaning becomes immediately apparent.

Searching an index for a listing of every time a certain item occurs can be revealing for an infrequently used word. But generally too much information comes back. If you expect a software program to find meaningful matches, there must be an association of items. If the association of those items is too broad then you have still failed to find relevant information.

Human intelligence depends upon its ability to discern differences, similarities, and identities. Thinking includes the grouping of commonalities into classes. It is easier to mentally manipulate broad categorizations than to keep track of a million details. These collections of similarities can be thought of as sets.

Language sets can be composed of many things: the set of associated words which describe a concept, the set of characters and digits which describe different part numbers, the set of sequenced characters and spaces which have been used to delineate a frequently misspelled word, the set of words, digits, and punctuation which delineate a range of values, and so on.

By locating sets in relation to other sets, relevance can be pinpointed. Therefore an intelligent text retrieval program must allow for specification of any type of set of lexical items, intersecting with other such sets, inside of meaningful communication units. It is upon this mandate that Metamorph was designed.

If we look at the word "bear" by itself and match every occurrence, a reference to "the brown bear in the woods" is as equally valid as "she just couldn't bear any more abuse". A search for "arms" by itself cannot be faulted for locating "his arms were long" even if you were more interested in the subject of firearms.

But if instead we look for "bear arms" within a sentence, it becomes a simple matter to locate "The Constitution clearly safeguards the right to bear arms." And since we are dealing with sets, it will be just as easy to locate "He hoped that today his father would let him carry the rifle". Where the concept "bear" (which includes "carry"), and the concept "arms" (which includes "rifle") cross inside the natural delimiters of a sentence, we have found relevance.

In the same precise manner we can poll an incoming newswire for a paragraph containing the set of possible spellings of "Larousche", the set of words associated with "election", and the set of word and digit combinations which have a numeric value between 1 and 100%, to see if anyone knows how Lyndon Larousche actually came out in the recent election.

In this fashion, degree of required relevance, how wide or how narrow the realm of possible matches, is entirely under the user's control. The statement of the query is interpreted as a number of sets, which must intersect (or not) within some naturally defined boundary.

Where these sets intersect is the researcher's target. The task of locating these points with speed and precision is Metamorph's job.

Metamorph Search Strategy

Metamorph reads text, just like a person reads text, in sequence. This is called a linear search, or a free text scan. Even where indexing is used for optimization, the linear read is always done as a final step.

A query is interpreted as a number of items. Each of these items can be expanded into a set of possible matches. Metamorph has different methods available for searching based on the type of item and its set. Its mission is to locate places in the text where these sets intersect.

Having read a big chunk of text into its buffer, Metamorph picks the most efficient way to locate at least one qualifying search item. Then it looks backwards and forwards to the beginning and ending delimiters being used, to see if there is another qualifying search item. And so it goes, until it finds a unit of text containing occurrences of everything required.

The delimiter boundaries are defined as expressions which the program understands. For paragraphs this might be two new line characters in a row. For sentences this could be the ending sentence punctuation. These expressions can vary within the scope of a definable regular expression.

Every action done is designed for optimum speed. Anything which does not contribute to qualifying whether a portion of text meets the requirements of the query is considered unnecessary.

A search which takes more work, like evaluating words in text as quantities, is rooted wherever possible to searches which are very fast, such as a word or string match. By rooting a search is meant that the fastest search is done first; then the more detailed search need only be done on the text within the delimiters surrounding the found item.

Using the 250,000+ word Thesaurus provided with Metamorph, most words can be expanded to word sets which can be quite large. The power of the search is not in the size of the set, but in the intersection of these sets. Where they cross lies meaning.

A Metamorph search for "acquisitions greater than 1 million" quickly reveals a buried reference to "a purchase offering was accepted last night for four and a half billion on behalf of Acme High Tech Inc." Or it takes a query like "Has there been a power struggle in the Near East?" and skips over unrelated references, presenting immediately "The conflict in Kuwait forced a confrontation."

"Power" is automatically expanded to a set of 57 possible words, including "force". "Struggle" is expanded to 23 associations, including "conflict". "Near East" becomes a set of 6 including "Kuwait". There may be hundreds of articles on the Near East, or even on power struggles, but there are only a few relevant responses containing an intersection of all 3 of these sets.

In the same sense, many specific or technical words do not require equivalent associations. The power of a technical search is not in finding its synonym, but in finding the exact technical term in any form, in association with some other set or sets of concepts.

Vocabulary is as personal as the people who write or speak. Therefore the user, rather than a detached programmer, can teach Metamorph what's important about one's own language usage. In so doing Metamorph's Thesaurus can evolve with every search, its updating an automatic by-product of continued use.

But synonyms are only part of the picture. Sometimes it is merely the presence of several one item sets that a person seeks. Software documentation is as often as not searched with the Thesaurus off. The ability to find the one line in the Unix manuals where "files", "copy", and "super user" intersect could save hours of wasted time and be make or break on the efficiency of an automated help desk.

In using Metamorph, one learns how best to state queries so that the desired associations in text can be found. The more experienced user can take advantage of learned syntax to locate its special search items. One can get quite complex in the statement of what is to be found, and in making use of a multitude of program features.

What is most important for satisfactory research results is remembrance of program goal; that is, to let Metamorph help you find relevant text fast, so you can use that information in the best way possible.

Text Metamorph can Search

Metamorph can search any file, preferring text of flat ASCII format. Metamorph can search files which are not flat ASCII, but it is the ASCII characters which will be recognized.

Where files contain a mixture of text and graphics, the graphics characters are skipped over; the ASCII text is recognized and retrieved. You can even search a binary file for text strings.

When text is found it is displayed with Metamorph's Browser, as it was entered. Certain control characters are filtered out for cleaner viewing.

Since Metamorph reads files in sequence from beginning to end, the usefulness of the responses is dependent upon content being stored in correct sequence. You can find ASCII strings in a database file; the retrieved text will make sense to the degree that string makes sense in its stored context.

The files to be searched need not be loaded, indexed, or prepared in any way; you simply specify the names of the files you wish to search from the Files Selection Menu. Metamorph reads those files where they are, as they are, without the need for any modification Multiple files can be searched in any combination across drives and directories

Metamorph can search stream data live as it comes off a wire or circuit. Therefore you can profile information as it is coming in off live wire data feeds without the necessity of downloading it first. Any query which can be executed against static information can also be run in batch in such a dynamic application

Definition of Terms

Query:

A Query is the question or statement of search items to be matched in the text. A Query is comprised of one or more search items, which can be of different types.

Hit:

A Hit is the text Metamorph retrieves in response to a query, whose meaning matches the Query to the degree specified.

Search Item:

A Search Item is an English word or a special expression. A word is automatically processed using certain linguistic rules. Special searches are signaled with a special character leading the item, and are governed respectively by the rules of the pattern matcher invoked.

Set:

A Set is the group of possible strings a pattern matcher will look for, as specified by the Search Item. A Set can be a list of words and word forms, a range of characters or quantities, or some other class of possible matches based on which pattern matcher Metamorph uses to process that item.

Intersection:

A portion of text where at least one member of two Sets is matched.

Delimiters:

Delimiters are repeating patterns in the text which define the bounds within which search items are found in proximity to each other. These patterns are specified as regular expressions.

Intersection Quantity:

The number of unions of sets existing within the specified Delimiters. The maximum number of Intersections possible for any given Query is the maximum number of designated Sets minus one.

Hits can have varying degrees of relevance based on the number of set intersections occurring within the delimited block of text, definition of proximity bounds, and weighting of search items for inclusion or exclusion.

Intersection quantity, Delimiter bounds, and Logic weighting can be adjusted by the user as part of Query specification.

Concept Set Intersection

Metamorph's vocabulary is around 250,000+ word connections, constructed in a dense web of associations. This vocabulary is stored in a Thesaurus, also called an Equivalence File. Proximity of concept can be fine tuned to qualify degree of relevance, providing matches which are sometimes concrete, sometimes abstract.

A keyword entered as a Search Item is looked up in the Thesaurus for associated words it can deem equivalent to the entered root word. This list of words and their word forms comprise the keyword's concept set. Text containing meaning relevant to your Query can be retrieved by locating places in the text where more than one set of concepts meet.

The content of the concept sets and the number of intersections present in a given hit determine how relevant the response will be to the stated query. The definition of proximity bounds as a sentence, paragraph or some other designated block of text can determine how tightly or loosely these concepts are correlated.

Default Operation of a Metamorph Search

A default Metamorph search calculates the maximum number of intersections possible in a hit based on your Query.

Metamorph picks out the important words in a question, and counts the number of sets. It eliminates noise, and does a search for any hit containing matches to the remaining items. Each valid search item is assigned equal weight.

If you ask: "Was there a power struggle?" the program looks for hits containing one intersection of the two sets "power" and "struggle".

Search items are sought within the proximity bounds of a sentence. A "sentence" is specified as the block of text from one sentence ending to the next sentence ending. Once the first search item is found, Metamorph looks for other qualifying search items backwards to the last sentence ending, and forward to the next.

When a qualifying hit is found, it is brought up in context of the full text in which it was located, and is available for browsing. Those words or strings of characters in the text which matched the search items are shown in contrast. The entire hit is highlighted.

The name of the text file in which the hit is found as well as the entered Query is always shown along with the hit.

Default settings are stored as variables in a profile. As such, almost every aspect of a search can be adjusted by the user through the User Interface.

Tailoring Metamorph's Linguistics

Concept sets can be edited by the user to include special vocabulary, acronyms, and slang. There is sufficient vocabulary intelligence off the shelf so that editing is not required to make good use of the program immediately upon installation. However, such customization is encouraged to keep online research in rapport with users' needs, especially as search routines and vocabulary evolve.

A word need not be "known" by Metamorph for it to be processed. The fact of a word having associations stored in the Thesaurus makes abstraction of concept possible, but is not required to match word forms. Such word stemming knowledge is inherent. And any string of characters can be matched exactly as entered.

You can edit the special word lists Metamorph uses to process English if you wish. As it may not be immediately apparent to what degree these word lists may affect general searching, it is cautioned that such editing be used sparingly and with the wisdom of experience. Even so, what Metamorph deems to be Noise, Prefixes, and Suffixes is all under user control.

Noise is defined as the small, common, relational words which appear frequently in a particular language and refine and fine tune specific communications, but do not majorly affect the larger concepts under discussion; e.g., about, in, on, whether. Pronouns, question words, and state of being verbs are treated as Noise for the purpose of a Metamorph search.

Suffixes are the common endings to words which modify tense and form but do not change the basic meaning of the word; e.g., -ing, -es, -ed, -tion, -ary. Prefixes are common syllables added to the beginning of a word which change the meaning in some way; e.g., re-, pre-, un-, dis-.

Any of these lists can be added to or edited at will, entirely at the user's discretion.

Types of Searches

Many types of searches are possible. If you do not otherwise specify, a word entered as part of a query is treated as English, and passed to a pattern matcher which follows certain English rules.

You can search for English words by themselves (that word and its word forms only), as a set of words and their word forms, or as part of a phrase.

You can also search for a literal string of characters, and can include wildcards (*) to fill in the parts you don't know, rooted to the portion of a string you do know.

You can call other special pattern matchers with a special character leading your expression. This enables the user to find fixed and variable length regular expressions, approximated expressions, or numeric quantities existing in text as English words, letters, and digits.

A special expression can be searched for in combination with another special expression of the same type, or with a different special expression, or with other English words. Any valid search item can be searched for as intersecting with another valid search item.

It is intended that what is easiest to enter, that is English words, be interpreted in a way which will get the most satisfying user results without the need for a long education in how to specify a Query. On this basis the program defaults have been derived. It is encouraged that the more demanding user study the information in the manual on how to construct a query, as well as the Supplements detailing special query syntax, to realize full functionality provided with the program.

Our program motto is "If it's there, we can find it." Use Metamorph's different search capabilities as a benchmark for accuracy regarding existence or absence of information in text.

Metamorph as a Hypertexting Tool

"Hypertexting" is making hooks from coarse information into more detailed or refined information, which can be selectively activated.

Indexing programs offer a fast search against easily indexable subjects, titles, and authors, but do not provide the ability to look for proximity of concepts within the narrative of such entries.

Using Metamorph as a first level operator on an overview file of abstracts or synopses, one can find discrete correlations of concept at a first pass. This is much like going to a card catalog in a library and being able to read the whole entry; Metamorph can do concept searching against the descriptive abstract as well as just the title, author, and subject. Once the right entry is located, a hook to some other information, image, or media type can be called up.

Using tools provided with the Metamorph package, it is possible to mark certain sections of text so that a first pass search is done on portions of the full text only. This makes it possible to retain a linear concept search against narrative text, while limiting the amount of information Metamorph must search through at a first pass.

Once a hit is found which represents an intersection of the concepts designated by the querier, the user can selectively launch another search on certain preselected data by pressing a hot key. The tagged data can be made to hook to other information by calling up another Metamorph on designated data files, or calling up associated figures or diagrams or tables or graphics displayed in their native format.

This model can be extended to a CD ROM jukebox, where the user moves stepwise out through mythical terabytes of information. This approach is ideal for multimedia environments such as huge photographic, graphics, or audio libraries, where Metamorph is used against descriptive catalog information as a narrative switchboard for selectively launching the appropriate application.

Other Metamorph Applications

Automated message handling environments are an excellent environment for Metamorph since Metamorph has the ability to read a live wire feed as the data comes in without the necessity of preprocessing the data in any way before it can be searched.

News profiling of articles which are "hot off the press" is another place where reading the data as it arrives can be crucial for time sensitive issues.

Such systems would be minimally labor intensive to maintain and expand, using tools provided with the Metamorph package.

Since Metamorph does not require any preprocessing or indexing of textfiles before searching, it becomes an ideal companion for immediately searching text which has been scanned in using OCR (Optical Character Recognition) technology.

For those people who have C Programmer capability, custom user interfaces which call the Metamorph API (Application Program Interface) at a C code level can be designed to make use of Search functions in an application environment.

One of the most obvious applications would be where some in place database or document management program is used to organize and selectively retrieve data subsets from very large systems, then uses the API to pass that data subset to an intelligent concept search accomplished by Metamorph.

One can also use 3DB, Metamorph's database program, along with Metamorph to maximize efficiency and intelligence of search regardless of size.

Liaison Offices In:

Paris, France

London, England

Tokyo, Japan

FOR INFORMATION PLEASE CALL OR WRITE:

REAL-WORLD INTELLIGENCE, INC.
P.O. BOX 3566
WASHINGTON, DC 20007
U.S.A.
Phone: 202 338-1237
FAX: 202 298-6529

* * *

FOR TECHNICAL OFFICE
AND LABORATORY:

REAL-WORLD INTELLIGENCE, INC.
P.O. BOX 839
CHESTERLAND, OHIO 44026
U.S.A.
Phone: 216 729-7612
FAX or LOOKOUT POINT Online Service 216 729-8419

FIRST INTERNATIONAL SYMPOSIUM: NATIONAL SECURITY & NATIONAL COMPETITIVENESS: OPEN SOURCE SOLUTIONS Proceedings, Volume II - Link Page

Previous	Words Are Not Enough
Next	Commercial Remote Sensing: Open Source Imagery intelligence

Return to Electronic Index Page